Eötvös Loránd University
Institute of Mathematics

Ph.D. thesis

Sampling and local algorithms in large graphs

Endre Csóka

Doctoral School: Mathematics
Director: Miklós Laczkovich, member of the Hungarian Academy of Sciences

Doctoral Program: Pure Mathematics
Director: András Szűcs, member of the Hungarian Academy of Sciences

Supervisor:
László Lovász, member of the Hungarian Academy of Sciences

Department of Computer Science, Eötvös Loránd University
December 2012

Contents

Acknowledgement

I am very grateful to my supervisor László Lovász, and to Miklós Abért for all their help and guidance, which were either very useful or indispensable for each of my results. Beyond these results, I also learned a lot about how to make a theory from nothing. I believe that this experience will be useful in all of my life. I am happy that I could learn something about the outstanding approach of László Lovász to mathematics.

I want to say thank you to the co-authors of the results, and to Gábor Elek, especially for his guidance and observations for the undecidability results in Chapter 2.

Chapter 1

Introduction

Very large graphs are present in almost all areas of the world. These appear in biological systems, e.g. the brain; in physics, e.g. the graph of the bonds between the molecules of a solid; furthermore, the internet, the traffic system, the electrical grid and social networks like the acquaintance graph of all people are also important very large graphs. In many cases, these graphs are not only huge but it is hopeless to get to know them precisely. However, we still have a chance to get to know some important properties of them.

In the beginning, the statistical analysis of very large graphs became popular in other areas of science, especially in statistical physics. They measured and measure the degree distribution of the graph, sometimes together with the correlation of the degrees or the density of triangles, and some other "local" data. Then the conclusions are made from generating large random graphs with these parameters, and measuring the properties of these graphs. They usually use heuristic algorithms for generating random graphs, which do not guarantee uniform randomness at all. However, scientists of these areas are very satisfied with the results. Understanding the background of this phenomenon was the main motivation for the mathematical theory of very large graphs.

The mathematical description of the question is, which graph properties and parameters can be estimated by a constant-size sampling. A graph parameter is estimable if for each $\varepsilon > 0$, there exists a constant-time sampling algorithm such that for each graph, this returns a value with an error at most ε from the parameter value of the graph, in expectation. There are two different models for sampling algorithms.

The definition presented by Oded Goldreich, Shafi Goldwasser and Dana Ron [27] is the following. We take a constant number of vertices uniformly at random (allowing multiplicities), and we take the induced subgraph on these nodes. This is the simplified but equivalently strong version of the definition that we can use the following two kind of steps in a constant number of times. One kind of step is choosing a uniform random node and the other one is asking about two nodes whether there is an edge between them. The limit theory for this sampling method was developed by László Lovász with Balázs Szegedy [44], and with Borgs, Chayes, T. Sós and Vesztergombi [9], [8]. This theory could answer the main questions like what the testable properties and parameters are, therefore, this topic is complete in some sense. We call it the theory of dense graphs.

While this topic itself is complete, it has connections with several other topics, and there is further research in these directions. The language it uses already turned out to be useful for some existing and new topics. As an example, in Chapter 8, we show a new conjecture and some partial results on it. But returning to the original motivation, this theory is useful only for dense graphs, that is graphs with $\Theta(n^2)$ edges; but unfortunately, most real-life graphs are

not so dense at all, therefore, the sampling method returns with an empty graph for almost sure.

However, there is another model by Oded Goldreich and Dana Ron [28], with the limit theory developed by Itai Benjamini and Oded Schramm [5]. It deals with bounded degree graphs (or, at least, graphs with $O(n)$ number of edges). This model fits much better to the real-life networks, but its mathematical theory turned out to be a much more difficult task. While there are several important results about it, this theory is far from being completed. Moreover, this includes algorithmically undecidable questions, as we will show it in Chapter 4. The larger part of my dissertation is about this theory, called the theory of sparse graphs.

Here, sampling means the following. We choose a constant number of vertices uniformly at random, and we take the constant-radius neighborhood of each. This is the simplified but equivalently strong version of the definition that we can use the following two kind of steps in a constant number of times. One kind of step is choosing a uniform random node and the other one is getting the list of neighbors of a node.

There are some other topics about sampling from large structures, such as permutations by Kohayakawa [35], partially ordered sets by Janson [37], abelian groups by Szegedy [54] and metric spaces by Gromov [29] and Elek [23]. As the theory of dense graphs is the first and the only complete theory, therefore, this provides useful observations and suggestions for the other theories. We show an overview of these results and its connections to the theory of sparse graphs.

Homomorphism numbers are a common tool for the two models. For two graphs F and G, the homomorphism number $hom(F,G)$ is the number of edge-preserving mappings $h\colon V(F) \to V(G)$. Formally,

$$hom(F,G) = \left|\left\{h\colon V(F) \to V(G) \;\middle|\; \forall (x,y) \in E(F)\colon \big(h(x),h(y)\big) \in E(G)\right\}\right|.$$

In the theory of dense graphs and in the theory of sparse graphs, sampling a graph G can be expressed by getting approximate values for $t(F,G) = hom(F,G)\big/\big|V(G)\big|^{|V(F)|}$ and $hom(F,G)\big/\big|V(G)\big|$, respectively, for a bounded number of graphs F. That is, the only difference is the way of normalizing.

Consider the space $\mathcal{G}$ of all isomorphism types of graphs. We put a topology $\mathcal{T}$ on it, to express the similarity of graphs with respect to the sampling. We define $\mathcal{T}$ as the coarsest topology in which the homomorphism densities $t(F,.)$ are continuous for all graphs F. In other words, a sequence of graphs $G_1, G_2, ...$ is convergent in $\mathcal{T}$ if for all graphs F, the sequence $t(F,G_n)$ is convergent.

Notice that if we multiply all nodes of a graph by the same number, then these two graphs are equivalent in this topology, therefore, we do not distinguish them. On the other hand, graphs which do not arise from the same graph by node-multiplication of the same graph, are not equivalent in the topology.

Symmetric measurable functions $[0,1]^2 \to [0,1]$ are called *graphon*s. Graphs on the set of points $\{0,1,...,n-1\}$ are represented by the graphon defined by

$$w(x,y) = \begin{cases} 1 \text{ if there is an edge between } \lfloor nx \rfloor \text{ and } \lfloor ny \rfloor \\ 0 \text{ otherwise.} \end{cases}$$

Sampling from a graphon means that we take a constant number of uniform random values from $[0,1]$, we take the submatrix according to the rows and columns at these values, and we take the graph defined by this matrix as adjacency matrix. Sampling from a graph provides the same distribution as samples from the graphon representing the graph.

Two graphons $w_1, w_2 \colon [0,1]^2 \to [0,1]$ are weakly isomorphic if there exist two measure-preserving transformations $\sigma_1, \sigma_2 \colon [0,1] \to [0,1]$ such that for almost all pairs $(x,y) \in [0,1]^2$, $w_1(\sigma_1 x, \sigma_1 y) = w_2(\sigma_2 x, \sigma_2 y)$. Weakly isomorphic graphons provide the same distributions of samples.

Lovász and Szegedy [44] proved that the closure of $(\mathcal{G}, \mathcal{T})$ can be represented by the space of graphons, up to weak isomorphism. They also showed that this space is compact, which has important consequences, for example in extremal combinatorics.

Define the cut metric on the space of graphons in the following way.

$$\delta_\square(w_1, w_2) = \inf_{\sigma_1,\sigma_2} \sup_{S,T} \int_{x\in S} \int_{y\in T} w_1(\sigma_1 x, \sigma_1 y) - w_2(\sigma_2 x, \sigma_2 y) \, \mathrm{d}y \, \mathrm{d}x,$$

where σ_1 and σ_2 are $[0,1] \to [0,1]$ measure-preserving transformations, and S and T are measurable subsets of $[0,1]$. Lovász and Szegedy [44] showed the inequality

$$\forall F \in \mathcal{G}, w_1, w_2 \in W : \quad t(F, w_1) - t(F, w_2) \le |E(F)| \cdot \delta_\square(w_1, w_2).$$

There is a much more difficult inequation about the other direction, and these together imply that the topology indicated by the cut metric $\delta_\square$ is $\mathcal{T}$. In other words, the sequence of graphs G_n is convergent if and only if this is a Cauchy-sequence with respect to $\delta_\square$.

Summarizing, we embedded the space of all graphs into a nice and usable compact metric space, which expresses the similarity of graphs according to the sampling. Therefore, a parameter is estimable if and only if it extends to the space of graphons continuously. This space contains only graphs and limits of convergent graph sequences, therefore, if a continuous extension exists, then this is unique.

To show the power of this theory by an example, we can say that a graph is quasirandom if and only if it is close to a constant graphon in the cut metric. In other words, a graph is close to the constant p graphon if and only if its sample distribution is close to the sample distribution from an Erdős–Rényi random graph with parameter p.

Consider now the sparse graphs. In graphs large enough, the sampling provides pairwise disjoint neighborhoods with probability tending to 1. Therefore, we modify the sampling method to the following simpler and asymptotically equivalent form. For constants r and n, we consider the distribution of (radius) r-neighborhoods of a uniform random node, and we take n random elements from this distribution. We call it a *sample*.

There are two kinds of limit objects for bounded degree graphs, both have advantages and disadvantages. One is the graphings, introduced by Elek [20] and Aldous and Lyons [1]. A graphing is given by a finite set of measure-preserving bijections on a measure space. The other limit object is the random rooted graphs. This latter one fits better to our purposes.

Which real graph parameters can be estimated by sampling is a central question of this theory. Some examples for these parameters are the number of the nodes in the largest independent set, or dominating set, or the size of the maximum matching; or the smallest number of edges that should be deleted to make the graph planar, or to separate the graphs into components of sizes at most half of the original size, all of them normalized by the number of nodes. Formally, graph parameter is a function $p \colon \mathcal{G} \to \mathbb{R}$, and *estimator* is a function mapping from samples to real numbers. We say that a parameter p is estimable if for all $\varepsilon > 0$, there exists an estimator such that for all graphs G, the output of the estimator on a random sample from G is at a distance at most ε from $p(G)$ in expectation.

The estimability of a parameter expresses that the parameter is determined by its neighborhood distribution. Let us see this formally. Consider the space of all distributions of finite

or infinite size bounded-degree graphs, equipped with the sigma-algebra generated by the (discrete) distributions on constant-radius neighborhoods. Let us call them *random rooted graphs*. For all finite graphs G, we assign the random rooted graph $H(G)$ as follows. We choose a node uniformly at random, and we take its component with this root. We say that a sequence of random rooted graphs is convergent if for all radius r, the (finite dimensional) distribution of the r-neighborhoods of the root converge. Denote this topological space by X. It is not hard to see that p is estimable if and only if there exists a continuous real function $\tilde{p} : X \to \mathbb{R}$ on the topological space that extends p, that is,

$$\forall G \in \mathcal{G}\colon \qquad p(G) = \tilde{p}\big(H(G)\big). \tag{1.1}$$

Now we are ready to show that some of the parameters mentioned are not testable. Consider first the smallest number of edges that should be deleted to separate the graphs into components of sizes at most half of the original size, divided by the number of nodes. For a d-regular expander graph sequence, this ratio tends to a positive number. However, if for each graph, we take the disjoint union of two copies of it, then this ratio is 0, because the graph already has two components of half of the original size. But the neighborhood distribution of the two sequences tend to the same random rooted graph: the d-regular infinite tree (as a distribution concentrated on this only graph). Therefore, testability would require $\tilde{p}$ at the d-regular infinite tree to be 0 and that positive number at the same time, which is a contradiction.

Another important and less obvious example is the size of the maximum independent set, divided by the number of nodes. On any d-regular bipartite graph on $2n$ nodes, this expected ratio is $1/2$. On the other hand, on a random d-regular graph on $2n$ nodes, this ratio tends to less than $6/13$ if $n \to \infty$, as shown by Béla Bollobás [7]. As these two graph sequences tend to the d-regular infinite tree, as well, therefore, the independent ratio is not testable either. However, many other parameters, such as the relative size of the maximum matching is testable. Furthermore, the relative size of the independent set is testable for some special classes of graphs, including planar graphs.

In fact, (1.1) remains true even if $\tilde{p}$ is defined only on the closure of the set of distributions corresponding to graphs $cl\big\{H(G)\colon G \in \mathcal{G}\big\}$, which is a much smaller subspace. For example, if the degree of the root of a random rooted graph is 3 with probability 1, but all of its neighbors have degree 4, then this cannot be obtained as the limit of random rooted graphs $H(G_n)$ assigned to finite graphs G_n.

Denote the degree bound by d. Consider a random rooted graph, and change the root to each of its neighbors with probability $1/d$, and with the remaining probability, keep the root at the original node. This provides another random rooted graph. If the two random rooted graphs are the same (in distribution and up to isomorphism), then we say that the random rooted graph is *unimodular*. The space of unimodular random rooted graphs is denoted by X_u.

The random rooted graphs assigned to finite graphs are unimodular. By the conjecture of David Aldous and Russell Lyons, the other direction is also true for the closure, that is, $X_u = cl\big\{H(G)\colon G \in \mathcal{G}\big\}$. Or equivalently, for all unimodular random rooted graphs U, there exists a sequence of graphs G_n such that the corresponding random rooted graphs $H(G_n)$ tend to U. This conjecture is already known in some special cases, e.g. when the distribution is concentrated on trees. [11] [21]

It turned out that the Aldous–Lyons Conjecture is strongly related to other important topics, as well. There are a number of conjectures for all countable discrete groups which are proven only for sofic groups. Mikhail Gromov asked whether all countable discrete groups are sofic. This was conjectured to be false, but there was no counterexample. Later, Gábor Elek showed, using the Cayley-graphs of the groups, that a version of the Aldous–Lyons Conjecture

would imply the positive answer for the question of Gromov, which, of course, would imply the positive answer for all those conjectures for all countable discrete groups.

Therefore, and also independently of this, it would be useful to describe the space of all unimodular random rooted graphs X_u, and the closure of the space of the random rooted graphs obtained from graphs $cl\{H(G)\colon G \in \mathcal{G}\}$. Unfortunately, this is hopeless, because these subspaces have no nice description. Namely, in Chapter 4, we will show that some natural questions about the shape of them are algorithmically undecidable. But we also mention that the answers to all these questions are the same for the two sets, in accordance with the conjecture.

In Chapter 2, we will show that if the Aldous–Lyons Conjecture is false, then there exists a unimodular random rooted graph that, with high probability, can be distinguished from the finite graphs by the constant-radius neighborhood of only one random node. As the tool for this proof, we show that an approximately maximum flow can be constructed by a deterministic local algorithm on bounded degree graphs. Local algorithm is a concept strongly related to parameter estimation, and defined in the next subsection.

Local algorithms

A distributed algorithm on bounded degree graphs means the following. We place a processor at each vertex of the input graph, and two processors can directly communicate if they are at neighboring nodes. At the end, each processor makes some decision, and this is the output of the algorithm. For example, if we want to find a large matching, then at the end, each processor decides which of its neighbors to match with, or whether to keep unmatched. Of course, these decisions should be consistent. Distributed algorithms can be defined in several nonequivalent ways.

A local algorithm is a distributed algorithm that runs in a constant number of synchronous communication rounds, independently of the number of nodes in the network. An equivalent definition of local algorithms is that the output of each node is a function of (the isomorphism type of) the constant-radius neighborhood of the node.

Research on local algorithms was pioneered by Angluin [3], Linial [42], and Naor and Stockmeyer [49]. Angluin [3] studied the limitations of anonymous networks without any unique identifiers. Linial [42] proved some negative results for the case where each node has a unique identifier. Naor and Stockmeyer [49] presented the first nontrivial positive results. For more about local algorithms, see the recent survey paper by Suomela [53].

Randomness is a powerful and classical technique in the design of distributed algorithms, and particularly useful in breaking the symmetry [2, 36, 46]. For example, on transitive graphs, any local algorithm should choose the same output at each node; therefore, it is impossible to choose a positive fraction of independent vertices by a deterministic local algorithm, but this is possible with randomization. An equivalent description of random local algorithms is the following. We assign independent random seeds to the nodes, and the output at each node depends only on the constant-radius neighborhood of it, including the random seeds assigned to the vertices in the neighborhood. For example, if we choose the nodes which have a higher seed than all their neighbors, then we get an independent set of expected relative size at least $1/(d+1)$.

For typical problems, we expect from local algorithms approximate solutions only. For example, we say that we can find an almost maximum independent set if for each $\varepsilon > 0$, there exists a local algorithm that for each graph G, outputs an independent set, and the expected size of this set is at most εn less than the size of the maximum independent set in G.

Local algorithms are strongly related to parameter estimation, because many of the examined parameters come from maximization problems. For example, the size of the maximum matching, the size of the maximum independent set, or the size of the maximum cut, normalized by the number of nodes. The connection is that if we have a random local algorithm which provides an almost optimal structure, e.g. an almost maximum matching, then the relative size of the maximum matching is an estimable parameter. The estimator is the following. We take the same radius neighborhoods of a constant number of random nodes, as the radius that the random local algorithm uses. For each neighborhood, we assign random seeds to the vertices and then we calculate whether the algorithm would match the root. Then the half of the ratio of the matched nodes gives a good approximation for the relative size of the maximum matching. Huy Ngoc Nguyen and Krzysztof Onak [50] proved the estimability of several problems, e.g. the relative size of the maximum matching, by creating local algorithms.

A local algorithm with preprocessing means that the output of each node is a function of (the isomorphism type of) the graph and the constant-radius neighborhood of the node. Or equivalently, each vertex receives the same "central information" depending on the entire graph and then they make a constant number of synchronous communication rounds, and then they present the output. This can also be interpreted as a service, as follows. There is a center with arbitrary information about the entire graph. Taking the maximum independent set problem as an example, each vertex can anonymously ask the center whether it is in the set, and the center should answer using its preprocessed information from the graph and (the isomorphism type of) the constant-radius neighborhood of the node. These answers must be consistent, namely no two neighboring nodes should receive "yes", but the proportion of nodes receiving "yes" should be close to the relative size of the maximum independent set.

Gábor Elek [20] proved the estimability of several parameters on graphs of subexponential growth, using a random local algorithm with the following preprocessing. He made a finite statistics of constant-radius neighborhoods of random nodes. In other words, the output at each vertex could depend on its constant-radius neighborhood and this statistics. The point of this concept is that the existence of such an algorithm giving good approximation still implies the estimability, as follows. We use the constant number of the same radius neighborhoods to make the statistics, we give this statistics to one further neighborhood, and we calculate the decision at the root. We repeat this procedure a constant number of times, from which we can get an estimation for the parameter. This observation was used for a tool to convert some statements in Borel graph theory to theorems in the field of local algorithms [24].

In Chapter 3, we compare the strengths of the different generalizations of local algorithms. In particular, we show that preprocession is useless. More precisely, if there exists a local algorithm using preprocession, then there exists another local algorithm with the same radius, in which the only "preprocession" is a random variable with a continuous distribution, and this provides an output with at most the same error from the optimum, in expectation. The use of this random variable can also be interpreted as follows. We draw a local algorithm from a given probability distribution of local algorithms, and we use this at each node.

Returning to the original motivation of the sampling method, there is an experience from physicists that everything is local in all real-life graphs as well as in random graphs. While this statement is very vague, physicists assume such statements for their results whenever they can use them, and they are satisfied with the results. We show an example for a result in Chapter 7, where we give an explanation for a phase transition, using an assumption about locality. The mathematical results provided this assumption coincide with the simulation results on random graphs.

We try to find a true mathematical statement expressing the experience that "everything is

local" on random graphs. In addition, this could open the door to describe the terms "typical" or "real-life" graphs better: a graph is as much "typical" as true that "everything is local" on it. There are many statements about graphs which are not true for all graphs, but which are true for the typical graphs. About such statements, all what we can do is proving that this is true for uniform random graphs on a large vertex set, with probability tending to 1. The theoretical imperfectness of this technique is pointed out by computer programs that are able to distinguish between uniform random graphs and different kind of real-life graphs, with high probability. Therefore, something to be true for almost all graphs does not mean that it is true for the typical graphs. The curiosity of the result in Chapter 7 is it proves something for typical graphs, but in an unusual sense.

We are still in the progress of describing the experience mathematically. In order to find the right statement, the best we can do is to elaborate on the easiest special cases.

The easiest problem of this kind is about the independence ratio of the 3-regular large girth graphs. First, this is a simple case because its neighborhood statistics is concentrated into the 3-regular tree. Second, the independence ratio is not determined by the neighborhood statistics. Third, it is clear what we mean random graph in this case, because the uniform random 3-regular graph has this neighborhood distribution with arbitrary small error, with probability tending to 1.

The expected relative size of the independent set generated by a local algorithm depends only on the statistics of the constant-radius neighborhoods of the graph (see Lemma 3.3). Furthermore, for a given neighborhood statistics, the supremum of this relative size by local algorithms is a lower bound for the independence ratio of graphs with this statistics. The experiences shown above suggest that we can construct an almost maximal independent set on a random graph. This conjecture can be split into two parts. One is that given the neighborhood statistics, the random graph has the lowest relative size of the maximum independent set. The other is that the lowest ratio is approximable by a local algorithm. Both statements would be somewhat surprising; therefore, this question is a good indicator of whether we are on a good track to understand the experiences shown above. In Chapter 5, we show a local algorithm providing the highest independence ratio achieved so far on 3-regular large girth graphs.

There is a natural limit of local algorithms, called factor of i.i.d. (independent identical distribution) processes. This means that we assign the random seeds to the vertices, and the "local" decision should be a measurable function of the rooted graph, in a natural sense. In this language, our previous question is about the largest independence ratio achievable on 3-regular infinite trees by factor of i.i.d. processes.

While approximately maximum matching is achievable by local algorithm, it is meaningful to ask whether there exists a factor of i.i.d. maximum matching. This question is also related to group theoretic questions, through their Cayley-graphs. To show another motivation, in the proof of Miklós Laczkovich [40] for Tarski's circle-squaring problem, the main idea was to find a perfect matching in a graphing defined by a finite number of translations. But it is still an open question whether the equidecomposability of the square and circle can be achieved with Lebesgue-measurable pieces. A similar construction with factor of i.i.d. perfect matching could provide a measuable equidecomposition.

Russell Lyons and Fedor Nazarov [47] proved that in every bipartite Cayley-graph of every non-amenable group, there is a factor of i.i.d. perfect matching. The nonbipartite case is much more difficult, but we prove in Chapter 6 that there is a factor of i.i.d. perfect matching in all nonamenable Cayley-graphs, or, in fact, in all nonamenable vertex-transitive unimodular random graphs.

Chapters 2, 3 and 4 are results of the author, based on the papers [16], [17] and [18],

respectively. Chapter 5 is a joint work with Balázs Gerencsér, Viktor Harangi and Bálint Virág, Chapter 6 is a joint work with Gábor Lippner [19], Chapter 7 is a joint work with Márton Pósfai and Yang-Yu Liu [43], and Chapter 8 is a joint work mainly with Tamás Hubai and László Lovász, but also with Omar Antolín Camarena and Gábor Lippner [14].

Chapter 2

Deterministic local algorithms for the maximum flow and minimum cut

A testable parameter should be asymptotically invariant under the modification of $o(n)$ nodes and edges, because with probability tending to 1, these do not modify the sample. In a network, if we delete all edges from all sources or targets, then the value of the maximum flow decreases to 0. Therefore, the value of the maximum flow in a graph with 1, or even with $o(n)$ sources or targets cannot be tested in any reasonable way. Similarly, one new edge with high capacity between a source and a target would increase the value of the maximum flow by this arbitrary large value. These are some reasons why we will deal only with multiple sources and targets and bounded capacities.

2.1 Model and results

There is an input network $N = (G, c)$, as follows. $G = (S, R, T, \mathbf{E})$ is a graph with degrees bounded by d. d is a global constant throughout the paper. The vertices of G are separated into the disjoint union of the sets S (source), R (regular) and T (target). $\mathbf{E}$ is the set of directed edges of G, which satisfies that $(a, b) \in \mathbf{E} \Leftrightarrow (b, a) \in \mathbf{E}$. We have a capacity function $c\colon \mathbf{E} \to [0, 1]$ of the directed edges. We will use the terms "graph", "path" and "edge" in the directed sense. Let $V = V(G) = V(N) = S \cup R \cup T$, $|V| = n$, $out(A) = \big\{(a, b) \in \mathbf{E} \;\big|\; a \in A, b \notin A\big\}$, and $out(v) = out\big(\{v\}\big)$, and for an edge $e = (a, b)$, let $-e = (b, a)$ denote the edge in the opposite direction.

A function $f\colon \mathbf{E} \to \mathbb{R}$ is called a *flow* if it satisfies the followings.

$$\forall e \in \mathbf{E}(G)\colon \qquad f(-e) = -f(e) \tag{2.1}$$

$$\forall e \in \mathbf{E}(G)\colon \qquad f(e) \le c(e) \tag{2.2}$$

$$\forall r \in R\colon \qquad \sum_{e \in out(r)} f(e) = 0. \tag{2.3}$$

The value of a flow f is

$$\|f\| = \sum_{e \in out(S)} f(e). \tag{2.4}$$

Denote a maximum flow by $f^* = f^*(N)$.

A set $S \subseteq X \subseteq S \cup R$ is called a *cut*. The value of a cut is

$$\|X\| = \|X\|_N = \sum_{e \in out(X)} c(e). \tag{2.5}$$

The Maximum Flow Minimum Cut Theorem [25] says that

$$\min_{S\subseteq X\subseteq S\cup R} \big\|X\big\| = \big\|f^*\big\|. \tag{2.6}$$

The rooted r-neighborhood of a vertex v or edge e, denoted by $h_r(v) = h_r(G, v)$ and $h_r(e)$, means the (vertex- or edge-)rooted induced subnetwork of the vertices at distance at most r from v or e, rooted at v or e, respectively. The set of all possible r-neighborhoods are denoted by $\mathcal{B}(r)$ and $\mathcal{B}^{(2)}(r)$, respectively. A function $F\colon \mathcal{B}^{(2)}(r) \to \mathbb{R}$ is called a *local flow algorithm*, and for each network N, we define the flow on N generated by F as $F(N) = \big(e \to F(h_r(e))\big)$. Similarly, $C\colon \mathcal{B}(r) \to \{true, false\}$ is called a *local cut algorithm*, and for each network N, we define the cut on N generated by C as $C(N) = \Big\{v \in V(G)\Big|C\big(h_r(v)\big)\Big\}$.

Theorem 2.1. *For each $\varepsilon > 0$, there exists a local flow algorithm F that for each network N,*

$$\big\|F(N)\big\| \geq \big\|f^*(N)\big\| - \varepsilon n. \tag{2.7}$$

Theorem 2.2. *For each $\varepsilon > 0$, there exists a probability distribution $\mathcal{D}$ of local cut algorithms such that for each network N, $F(N)$ is a flow, and*

$$\mathsf{E}_{C\in\mathcal{D}}\big\|C(N)\big\| \leq \big\|f^*(N)\big\| + \varepsilon n.$$

This implies that if we are allowed to generate a random seed (say, a sufficiently long string of 0-s and 1-s) and communicate it to every processor, then a cut can be computed whose capacity is almost minimum with high probability.

We mention that the same idea with much more accurate calculation than we will use would give an algorithm for each problem using radius $d^{1/\varepsilon}$ (if ε is small enough).

We also mention that we will prove in Theorem 3.12 that Theorem 2.2 is not true with any fixed local cut algorithm instead of a distribution of algorithms, not even if the local algorithm is allowed to use local random seeds (instead of our global random seed). In other words, there exists a constant real gap $g > 0$ that for each local cut algorithm C, there exists a network N that $\big\|C(N)\big\| \geq \big\|f^*(N)\big\| + gn$.

Corollary 2.3. *$\big\|f^*(N)\big\|/n$ is testable. In other words, for every $\varepsilon > 0$, there exist $k, r \in \mathbb{N}$ and a function $g\colon \mathcal{B}(r)^k \to \mathbb{R}$ such that if the vertices $v_1, v_2, \ldots v_k$ are chosen independently with uniform distribution, then*

$$\mathsf{E}\left(\left|\frac{\big\|f^*(N)\big\|}{n} - g\big(h_r(v_1), h_r(v_2), \ldots\ h_r(v_n)\big)\right|\right) < \varepsilon.$$

We note that for all $\varepsilon > 0$, having an expected error less than ε is a stronger requirement than having less than ε error with at least $1 - \varepsilon$ probability; but these are equivalent if the error is bounded.

2.2 Proofs

First we prove Theorem 2.1 using the following lemmas, and Corollary 2.3 will be an easy consequence of it.

An augmenting path of a flow f is a directed path $u = (e_1, e_2, \ldots e_k)$ from S to T with $f(e_i) < c(e_i)$ for each edge e_i. The capacity of u means $cap(u) = cap(u, f) = \min_i(c(e_i) - f(e_i))$, and we identify an augmenting path u with the flow $u\colon \mathbf{E}(G) \to \mathbb{R}$, $u(e) = \{1$ if $\exists i\colon e = e_i$; -1 if $\exists i\colon e = -e_i$; 0 otherwise$\}$, which we also call a *path-flow*. Augmenting on such a path u means the incrementation of f by $cap(u) \cdot u$.

Lemma 2.4. *If for a flow f, there is no augmenting path of length at most l, then*

$$\|f\| \geq \|f^*\| - \frac{d}{l}n. \tag{2.8}$$

Proof. $f^* - f$ is a flow on the network $(G, 2)$ (graph G with identically 2 capacity function). Therefore this can be decomposed into the sum of path-flows $u_1, u_2, \dots u_q$ and a circulation u_0 that follow the directions of the flow $f^* - f$, namely, for every $i \in \{0, \dots q\}$ and $e \in \mathbf{E}(G)$, we have $sgn\big(u_i(e)\big) \in \big\{0, sgn\big((f^* - f)(e)\big)\big\}$. For example, we can do it by the Ford-Fulkerson algorithm [25] on the network $\big((S, R, T, \{e \in \mathbf{E}(G) \colon (f^* - f)(e) > 0\}), 2\big)$. Thus,

$$\|f^*\| - \|f\| = \|f^* - f\| = \sum_{i=1}^{q} \|u_i\| = \frac{1}{2l} \sum_{i=1}^{q} 2l\|u_i\|$$

$$\leq \frac{1}{2l} \sum_{i=1}^{q} \sum_{e \in \mathbf{E}(G)} u_i(e) = \frac{1}{2l} \sum_{e \in \mathbf{E}(G)} \sum_{i=1}^{q} u_i(e) \leq \frac{1}{2l} \sum_{e \in \mathbf{E}(G)} 2 \leq \frac{d}{l}n. \qquad \square$$

Lemma 2.5. *If for a flow f, there is no augmenting path shorter than k, then augmenting on a path of length k does not create a new augmenting path of length at most k.*

Proof. Let the residual graph of a network $N = (G, c)$ with respect to a flow f be the graph $G_f = \Big(S(G), R(G), T(G), \big\{e \in \mathbf{E}(G) \big| f(e) < c(e)\big\}\Big)$. Then the augmenting paths of G can be identified with the paths in G_f from S to T. So if the length of the shortest augmenting path of a flow is k, then it means that the length of the shortest path in G_f from S to T is k. Let the movement of an edge of N mean the difference of the distances of its endpoint and starting point from S in G_f. Augmenting on a shortest path adds only such edges to the residual graph on which f decreases, which are the reverse edges of the path. All these edges have movement -1 (calculated before augmenting). So if a path becomes an augmenting path at this augmenting step, then all its edges have movements at most 1 and contain an edge with movement -1, so its length is at least $k + 2$. $\square$

Let us fix l, and call the paths of length at most l "**short paths**". Let us take a uniform random order σ on all short paths among the orders satisfying that shorter paths always precede longer paths. (This starts with the 1-lenght paths in uniform random order, than the 2-length paths in uniform random order, ..., the l-length paths in uniform random order.) We define the **chain** as a sequence $u_1, u_2, \dots u_s$ of short paths which is in the reverse order of σ, and satisfies that $\forall i \in \{1, 2, \dots s - 1\}$ there exists a common undirected edge of u_i and u_{i+1} (henceforth: these *intersect* each other).

Lemma 2.6. *For each $l \in \mathbb{N}$ and $\varepsilon > 0$ there exists $q = q(l, \varepsilon) \in \mathbb{N}$ that for every graph G (with degrees bounded by d) and its undirected edge e, with random order of all short paths, the probability that there exists a chain $u_1, u_2, \dots u_q$ for which u_1 contains e is at most ε.*

Proof. There exists an upper bound $z = z(l)$ for the number of short paths that intersect a given short path. Hence, there are at most z^q sequences of short paths $u_1, u_2, \dots u_q$ for which e is in u_1 and $\forall i \in \{1, 2, \dots q - 1\}$, u_i intersects u_{i+1}. All such sequences contain $\lceil q/l \rceil$ paths of the same length. Their order is chosen uniformly among the $\lceil q/l \rceil!$ permutations, so the probability that these are in reverse order is $1/\lceil q/l \rceil!$. This event is necessary for the sequence to be a chain. Denote the number of chains as in the lemma by the random variable X (as a function of the random order). We have

$$P(X \geq 1) \leq \mathsf{E}(X) \leq \frac{z^q}{\lceil \frac{q}{l} \rceil !} \to 0 \text{ where } q \to \infty,$$

which proves the lemma for some large enough number q. □

Proof of Theorem 2.1. Consider the variant of the Edmonds–Karp algorithm where we make the augmentations in the order σ, and we stop when no augmenting path of length at most l remains. In other words, we start from the empty flow, we take all short paths u in the order of σ, and with each path, we increase the actual flow f by $cap(f, u) \cdot u$. We denote this algorithm by A_1 and the resulting flow by $f_1 = f_1(N, \sigma)$.

Consider now the variant of the previous algorithm where we skip augmenting on each path which can be obtained as the first element of any chain of length s. We denote this algorithm by A_2 and the resulting flow by $f_2 = f_2(N, \sigma)$. The next lemma shows that $f_2(e)$ is a local algorithm.

Lemma 2.7. *For each edge e and order σ, the value of the flow f_2 on each edge e depends only on the sl-neighborhood of e, or formally,*

$$f_2(N, \sigma)(e) = f_2\big(h_{sl}(e), (\sigma|_{V(h_{sl}(e))})\big)(e).$$

Proof. Let us consider the execution of the two algorithms in parallel as follows. When the first algorithm takes a path u in G, then if u is in $h_{sl}(e)$, then the second algorithm takes u as well, otherwise it does nothing. If at a point, the two flows differ at an edge $\tilde{e} \in \mathbf{E}(h_{sl}(e))$, then there must have been a path u through $\tilde{e}$ on which the two algorithms augmented by different values. There are three possible reasons of it:

1. u is not in $h_{sl}(e)$;
2. u can be obtained as the first term of some chain in G of length s, but not in $h_{sl}(e)$;
3. u has an edge e' in $h_{sl}(e)$ at which the values of the two flows were different before taking u.

Assume that at the end, the two flows are different on e. Using the previous observation initially with $\tilde{e} = e$, let us take a path u through $\tilde{e}$ on which the two augmentations were different, and consider which of the three reasons occurred. As long as the third one, repeat the step with choosing $\tilde{e}$ as the e' of the previous step. Since by each step we jump to an earlier point of the executions, we must get another reason sooner or later. Denote the considered paths during the process by $u_1, u_2, \ldots u_t$. (Note that these are in reverse order on the augmenting timeline.)

Consider the case when the reason for u_t was the first reason. The set of all edges of all of these t paths is connected, it contains at most tl edges, it contains e and an edge at least sl away from e, so $tl > sl$, whence $t > s$. Thus $u_1, u_2, \ldots u_s$ is a chain with a connected edge set with size at most sl, so this chain is in $h_{sl}(e)$, therefore neither executions should have been augmented on u_1, contradicting the definition of u_1.

On the other hand, consider the case when the reason for u_t was the second reason. If we append $u_1, u_2, \ldots u_t$ with the chain from u_t and of length s, then as its subchain, we get a chain starting with u_1 and of length s, and it provides the same contradiction. □

We prove that if f_2 is the output of A_2 with $l = 2d/\varepsilon$ and using the function q of Lemma 2.6 and with $s = q(l, \varepsilon/(2d))$, then we have the following inequality.

$$\mathsf{E}\big(\|f_2\|\big) \geq \mathsf{E}\big(\|f_1\|\big) - \frac{\varepsilon}{2}n \overset{(2.8)}{\geq} \|f^*\| - \varepsilon n \tag{2.9}$$

First we prove the second inequality of (2.9). f_1 contains no short augmenting path, so using Lemma 2.4,

$$\|f_1\| \overset{(2.8)}{\geq} \|f^*\| - \frac{d}{l}n = \|f^*\| - \frac{d}{\frac{2d}{\varepsilon}}n = \|f^*\| - \frac{\varepsilon}{2}n.$$

To prove the first inequality of (2.9), we need the following lemma.

Lemma 2.8. *If $f_1(\sigma)(e) \neq f_2(\sigma)(e)$, then there exists a path through e which is the first term of a chain of length s.*

Proof. Let us consider the executions of A_1 and A_2 in parallel so that at the same time these take the same edge. If at a point of the executions, the two flows differ in an edge $\tilde{e}$, then there must have been a path u through $\tilde{e}$ on which the two algorithms augmented by different values. There are two possible reasons of it:

1. u is the first term of a chain of length s;
2. u has an edge e' on which the values of the two flows were different before taking u.

Assume that at the end, the two flows are different at e. Using the previous observation, let us take a path u through $\tilde{e}$ on which the two augmentation were different, and consider which of the two reasons occurred. As long as the latter one, repeat the step with choosing $\tilde{e}$ as the e' of the previous step. Since by each step we jump to an earlier point of the executions, we must get the first reason in finite many steps. Denote the paths considered during the process by $u_1, u_2, \ldots\ u_t$. Then appending $u_1, u_2, \ldots\ u_t$ with the chain from u_t and of length s, as its subchain, we get a chain of length s, starting with u_1. □

If $f_1(\sigma)(e) \neq f_2(\sigma)(e)$, then by Lemma 2.8, there exists a chain of length $q(l, \frac{\varepsilon}{2d})$, and Lemma 2.6 says that this has probability at most $\frac{\varepsilon}{2d}$. But even if this occurs, $f_1(\sigma) - f_2(\sigma) \leq 1 - (-1) = 2$. Therefore,

$$\mathsf{E}\big(\|f_1\|\big) - \mathsf{E}\big(\|f_2\|\big) = \mathsf{E}\big(\|f_1(\sigma)\| - \|f_2(\sigma)\|\big) = \mathsf{E}\big(\|(f_1(\sigma) - f_2(\sigma)\|\big)$$

$$= \mathsf{E}\Big(\sum_{e \in out(S)} \big(f_1(\sigma)(e) - f_2(\sigma)(e)\big)\Big) \leq \sum_{e \in out(S)} \frac{\varepsilon}{2d} \cdot 2 \leq \frac{d}{2}n \cdot \frac{\varepsilon}{2d} \cdot 2 \leq \frac{\varepsilon}{2}n,$$

which finishes the proof of (2.9).

Now, let $\bar{f}_2(e) = \mathsf{E}(f_2(\sigma)(e))$. It is a flow because it is easy to check that it satisfies all requirements, and $\|\bar{f}_2\| = \mathsf{E}(\|f_2\|) \geq \|f^*\| - \varepsilon n$. Furthermore, $\bar{f}_2(e)$ depends only on $h_{sl}(e)$, so it can be calculated by a local algorithm. Consequently, this algorithm satisfies the requirements of Theorem 2.1. □

Proof of Corollary 2.3. Let $\bar{f}_2$ denote the flow constructed by the local algorithm of the previous proof with error bound $\varepsilon/2$, which therefore satisfies $\|\bar{f}_2\| \in \big[\|f^*\| - \frac{\varepsilon}{2}, \|f^*\|\big]$, and let r be the radius used there plus 1. Using the notion $I(b) = \{1$ if b is true, 0 if false$\}$ for an event b, let

$$g\Big(h_r(v_1), h_r(v_2), \ldots\ h_r(v_k)\Big) = \frac{1}{k}\sum_{i=1}^{k} \Big(I(v_i \in S) \sum_{e \in out(v_i)} \bar{f}_2(e)\Big). \tag{2.10}$$

As $I(v \in S)\sum_{e\in out(v)} \bar{f}_2(e) \in [-d, d]$, the variance of (2.10) is at most $d/\sqrt{k}$, therefore (2.10) stochastically converges to its expected value, with respect to k. This expected value is

$$\frac{1}{n}\sum_{v\in V(G)}\Big(I(v\in S)\sum_{e\in out(v)}\bar{f}_2(e)\Big) = \frac{1}{n}\sum_{e\in out(S)}\bar{f}_2(e) = \big\|\bar{f}_2\big\| \in \Big[\big\|f^*\big\| - \frac{\varepsilon}{2},\ \big\|f^*\big\|\Big].$$

This implies that, for large enough k, these k, r and g satisfy the requirements. □

We define a **fractional cut** of a network N as a function $\tilde{X}\colon V(N) \to [0,1]$ so that

$$\forall s \in S\colon \tilde{X}(s) = 0 \text{ and } \forall t \in T\colon \tilde{X}(t) = 1. \tag{2.11}$$

The value of a fractional cut is defined as

$$\big\|\tilde{X}\big\| = \big\|\tilde{X}\big\|_N = \sum_{(a,b)\in\mathbf{E}(N)} c(a,b)\cdot\max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big). \tag{2.12}$$

Notice the following facts.

1. If X is a cut, then

$$\tilde{X}(v) = \begin{cases} 1 & \text{if } v \in X \\ 0 & \text{if } v \notin X, \end{cases} \tag{2.13}$$

is a fractional cut with

$$\big\|\tilde{X}\big\| \overset{(2.12)}{=} \sum_{(a,b)\in\mathbf{E}(G)} c(a,b)\cdot\max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \overset{(2.13)}{=} \sum_{(a,b)\in out(X)} c(a,b) \overset{(2.5)}{=} \big\|X\big\|.$$

2. If $\tilde{X}$ is a fractional cut, then for all $u \in (0,1)$, $X = \tilde{X}[u] = \big\{v \in V \big| \tilde{X}(v) < u\big\}$ is a cut, and for a uniform random u from $(0,1)$,

$$\mathsf{E}_u\Big(\big\|\tilde{X}[u]\big\|\Big) \overset{(2.5)}{=} \mathsf{E}_u\Big(\sum_{e\in out(\tilde{X}[u])} c(e)\Big) = \sum_{e\in\mathbf{E}} \mathbb{P}\Big(e \in out\big(\tilde{X}[u]\big)\Big)c(e)$$

$$= \sum_{(a,b)\in\mathbf{E}} \mathbb{P}(a < u \le b)c(e) = \sum_{(a,b)\in\mathbf{E}} c(a,b)\cdot\max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \overset{(2.12)}{=} \big\|\tilde{X}\big\|. \tag{2.14}$$

These calculations imply that $\min_X \big\|X\big\| = \min_{\tilde{X}} \big\|\tilde{X}\big\|$.

A map $\tilde{C}\colon \mathcal{B}(r) \to [0,1]$ is called a *local fractional cut algorithm*, and the fractional cut $\tilde{C}(N)$ produced by $\tilde{C}$ on a network N is the function $v \to \tilde{C}\big(h_r(v)\big)$.

Theorem 2.9. *For each $\varepsilon > 0$ there exists a local fractional cut algorithm that, for each network G, produces a fractional cut with value $\le \|f^*\| + \varepsilon n$.*

Proof. Let us execute the local flow algorithm as in Theorem 2.1 with error bound ε_1. Denote the resulting flow by f.

For a fix $\varepsilon_2 > 0$, we define the ε_2-residual graph as

$$G_f(\varepsilon_2) = \Big(S(G), R(G), T(G), \big\{e \in \mathbf{E}(G)\big| f(e) < c(e) - \varepsilon_2\big\}\Big). \tag{2.15}$$

Let $l = \lceil 1/\varepsilon_2 \rceil$. Denote the length of the shortest path from S to a vertex $v \in V$ in the graph $\big(G_f(\varepsilon_2)\big)$ by $dist(v)$. (If there is no path, then $dist(v) = \infty$.) We define the fractional cut $\tilde{X}$ as

$$\tilde{X}(v) = \begin{cases} \min\big(dist(v) \cdot \varepsilon_2,\ 1\big) & \text{if } v \in S \cup R \\ 1 & \text{if } v \in T. \end{cases} \tag{2.16}$$

This is a local fractional cut algorithm, because we need to execute the local flow algorithm for each edge only in the l-neighborhood of v, and then we get $\tilde{X}(v)$.

Let us focus on $\|\tilde{X}\|$. Notice that

$$\begin{aligned} \big\|f\big\| &\overset{(2.4)}{=} \sum_{e \in out(S)} f(e) \overset{(2.3)}{=} \sum_{e \in out(S)} f(e) + \sum_{r \in R} \Big(\big(1 - \tilde{X}(r)\big) \sum_{e \in out(r)} f(r)\Big) \\ &= \sum_{s \in S} \Big(1 \cdot \sum_{e \in out(s)} f(e)\Big) + \sum_{r \in R} \Big(\big(1 - \tilde{X}(r)\big) \sum_{e \in out(r)} f(r)\Big) + \sum_{t \in T} \Big(0 \cdot \sum_{e \in out(t)} f(e)\Big) \\ &\overset{(2.11)}{=} \sum_{v \in V} \Big(\big(1 - \tilde{X}(v)\big) \sum_{e \in out(v)} f(e)\Big) = \sum_{(a,b) \in \mathbf{E}} \big(1 - \tilde{X}(a)\big) f(a,b) \\ &\overset{(2.1)}{=} \sum_{f(a,b) > 0} \Big(\big(1 - \tilde{X}(a)\big) f(a,b) - \big(1 - \tilde{X}(b)\big) f(a,b)\Big) = \sum_{f(a,b) > 0} f(a,b)\big(\tilde{X}(b) - \tilde{X}(a)\big) \\ &= \sum_{f(a,b) > 0} f(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) - \sum_{f(a,b) > 0} f(a,b) \cdot \max\big(0,\ \tilde{X}(a) - \tilde{X}(b)\big) \\ &\overset{(2.12)}{=} \big\|\tilde{X}\big\|_{(G,f)} - \sum_{f(a,b) > 0} f(a,b) \cdot \max\big(0,\ \tilde{X}(a) - \tilde{X}(b)\big). \end{aligned} \tag{2.17}$$

Notice that the value of a fractional cut $\|\tilde{X}\|_{(G,c)}$ is additive as a function of c, because

$$\begin{aligned} \big\|\tilde{X}\big\|_{(G,c_1+c_2)} &\overset{(2.12)}{=} \sum_{(a,b) \in \mathbf{E}(G)} (c_1 + c_2)(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \\ &= \sum_{(a,b) \in \mathbf{E}(G)} c_1(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) + \sum_{(a,b) \in \mathbf{E}(G)} c_2(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \\ &\overset{(2.12)}{=} \big\|\tilde{X}\big\|_{(G,c_1)} + \big\|\tilde{X}\big\|_{(G,c_2)}. \end{aligned} \tag{2.18}$$

Using this observation with $c = f + (c - f)$, and using (2.17), we get

$$\begin{aligned} \big\|\tilde{X}\big\| = \big\|\tilde{X}\big\|_{(G,c)} &\overset{(2.18)}{=} \big\|\tilde{X}\big\|_{(G,f)} + \big\|\tilde{X}\big\|_{(G,c-f)} \\ &\overset{(2.17)}{=} \big\|f\big\| + \sum_{f(a,b) > 0} f(a,b) \cdot \max\big(0,\ \tilde{X}(a) - \tilde{X}(b)\big) + \big\|\tilde{X}\big\|_{(G,c-f)}. \end{aligned} \tag{2.19}$$

We call an edge $(a,b) \in \mathbf{E}(N)$ **bad** if there exists a short augmenting path in $G_f(\varepsilon_2)$ ending with the edge (a,b). Or equivalently, $(a,b) \in \mathbf{E}(N)$ is bad if $b \in T$ and there exists a path in $G_f(\varepsilon_2)$ from S to a of length at most $l-1$. Consider the terms $f(a,b) \cdot \max\big(0,\ \bar{X}(a) - \bar{X}(b)\big)$. If $f(a,b) \geq \varepsilon_2$, then $(b,a) \in \mathbf{E}\big(G_f(\varepsilon_2)\big)$ because $f(b,a) = -f(a,b) < -\varepsilon_2 \leq c(b,a) - \varepsilon_2$. By

(2.16), unless if (a,b) is bad, this implies $\tilde{X}(a) - \tilde{X}(b) \le \varepsilon_2$. Therefore if (a,b) is not bad, then either $f(a,b)$ or $\max\big(0, \bar{X}(a) - \bar{X}(b)\big)$ is at most ε_2, and both are $\in [0,1]$, therefore

$$f(a,b) \cdot \max\big(0,\ \bar{X}(a) - \bar{X}(b)\big) \le \varepsilon_2. \tag{2.20}$$

Since $\|f\| \ge \|f^*\| - \varepsilon_1 n$, there exist at most $\frac{\varepsilon_1}{\varepsilon_2} n$ edge-disjoint augmenting paths with capacity at least ε_2. Each short augmenting path intersect at most $z = z(l)$ short augmenting paths, which implies that there are at most $\frac{\varepsilon_1 z}{\varepsilon_2} n$ short augmenting paths with capacity at least ε_2. Therefore there exist at most $\frac{\varepsilon_1 z}{\varepsilon_2} n$ bad edges. Summarizing this and (2.20) and that the graph has at most $dn/2$ edges, we get

$$\sum_{f(a,b)>0} f(a,b) \cdot \max\big(0,\ \tilde{X}(a) - \tilde{X}(b)\big) \le \frac{\varepsilon_1 z}{\varepsilon_2} n + \frac{\varepsilon_2 d}{2} n \tag{2.21}$$

Consider now the terms of the sum

$$\big\|\tilde{X}\big\|_{(G,c-f)} \overset{(2.12)}{=} \sum_{(a,b)\in \mathbf{E}(N)} (c-f)(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \tag{2.22}$$

Assume that $(a,b) \in \mathbf{E}(N)$ is not bad. If $(c-f)(a,b) > \varepsilon_2$, then (2.15) shows that $(a,b) \in \mathbf{E}\big(G_f(\varepsilon_2)\big)$, therefore (2.16) implies $\tilde{X}(b) - \tilde{X}(a) \le \varepsilon_2$. Hence, similarly to (2.20), we get that $(c-f)(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \le \varepsilon_2$, therefore similarly to (2.21), we get

$$\sum_{(a,b)\in \mathbf{E}(N)} (c-f)(a,b) \cdot \max\big(0,\ \tilde{X}(b) - \tilde{X}(a)\big) \le \frac{\varepsilon_1 z}{\varepsilon_2} n + \frac{\varepsilon_2 d}{2} n. \tag{2.23}$$

Using $\varepsilon_1 = \varepsilon^2/(8dzn)$ and $\varepsilon_2 = \varepsilon/(2dn)$, we finish the proof by

$$\big\|\tilde{X}\big\| \overset{(2.12)}{=} \big\|f\big\| + \sum_{f(a,b)>0} f(a,b) \cdot \max\big(0,\ \tilde{X}(a) - \tilde{X}(b)\big) + \big\|\tilde{X}\big\|_{(G,c-f)}$$

$$\overset{(2.21,2.23)}{\le} \big\|f^*\big\| + 2\Big(\frac{\varepsilon_1 z}{\varepsilon_2} n + \frac{\varepsilon_2 d}{2} n\Big) = \big\|f^*\big\| + 2\Big(\frac{\frac{\varepsilon^2}{8dzn} z}{\frac{\varepsilon}{2dn}} n + \frac{\frac{\varepsilon}{2dn} d}{2} n\Big) = \big\|f^*\big\| + \varepsilon n. \qquad \square$$

Proof of Theorem 2.2. Consider the local fractional cut algorithm $\tilde{C}$ as in Theorem 2.9. We know that for each network N, $\big\|\tilde{C}(N)\big\| \le \big\|f^*\big\| + \varepsilon n$. For all $u \in [0,1]$, $\tilde{X}[u]$ is a local cut algorithm, and with a uniform random $u \in U[0,1]$, (2.14) shows that we get a probability distribution of local cut algorithms producing the desired expected value. $\square$

2.3 Applications on distributions of neighborhoods

Let $\mathcal{G}$ denote the set of all graphs (with degrees bounded by d). Let $H_r(G)$ denote the distribution of the r-neighborhood of a random vertex of a graph $G \in \mathcal{G}$. We call a family $\mathcal{F}$ of graphs **nice** if it is union-closed and closed under taking induced subgraphs (excluding the empty graph), that is, $G_1, G_2 \in \mathcal{F} \Rightarrow G_1 \cup G_2 \in \mathcal{F}$ and $G_1 \subseteq G \in \mathcal{F} \Rightarrow G_1 \in \mathcal{F}$ – where $\subseteq$ denotes nonempty induced subgraph – and $\emptyset \notin \mathcal{F}$. Let us denote the closure of the set of all r-neighborhood distributions in F by

$$D(\mathcal{F}, r) = cl\big\{H_r(G)\big|G \in \mathcal{F}\big\}. \tag{2.24}$$

For any two graphs $G_1, G_2 \in \mathcal{F}$ and $k, l \in \mathbb{N}$,

$$H_r\big((k \times G_1) \cup (l \times G_2)\big) = \frac{k|V(G_1)|H_r(G_1) + l|V(G_2)|H_r(G_2)}{k|V(G_1)| + l|V(G_2)|},$$

where $k \times G$ denotes the disjoint union of k isomorphic copies of G. This implies that each convex combination of the r-neighborhood statistics of two graphs in $\mathcal{F}$ can be approximated by the r-neighborhood statistics of another graph in $\mathcal{F}$. Therefore $D(\mathcal{F}, r)$ is a convex compact subset of $\mathbb{R}^{\mathcal{B}(r)}$. This implies that $D(\mathcal{F}, r)$ is determined by its dual, as follows.

Let us identify the natural base of $\mathbb{R}^{\mathcal{B}(r)}$ by the elements of $\mathcal{B}(r)$, and let the linear extension of $w\colon \mathcal{B}(r) \to \mathbb{R}$ defined as the function $\tilde{w}\colon \mathbb{R}^{\mathcal{B}(r)} \to \mathbb{R}$,

$$\tilde{w}\Big(\sum_{b \in B(r)} \lambda_b b\Big) = \sum_{b \in B(r)} \lambda_b w(b). \tag{2.25}$$

Let us define

$$m(\mathcal{F}, w) = \max_{P \in D(\mathcal{F},r)} \sum_{b \in \mathcal{B}(r)} w(b)P(b) \overset{(2.25)}{=} \max_{P \in D(\mathcal{F},r)} \tilde{w}(P) \overset{(2.24)}{=} \sup_{G \in \mathcal{F}} \tilde{w}\big(H_r(G)\big). \tag{2.26}$$

$D(\mathcal{F}, r)$ is determined by the values of $m(\mathcal{F}, w)$ for each $w\colon \mathcal{B}(r) \to \mathbb{R}$. Furthermore, for each $\lambda > 0$ and $c \in \mathbb{R}$, we have $m(\mathcal{F}, \lambda w + c) = \lambda m(\mathcal{F}, w) + c$. This shows that if we know $m(\mathcal{F}, w)$ for all $0 \le w \le 1$, then we know it for all w. Summarizing, for a distribution P on $\mathcal{B}(r)$,

$$P \in D(\mathcal{F}, r) \quad \Leftrightarrow \quad \forall w\colon \mathcal{B}(r) \to [0,1],\ \tilde{w}(P) \le m(\mathcal{F}, w).$$

We note without proof that the L_1-distance of a point $P \in \mathbb{R}^{\mathcal{B}(r)}$ from $D(\mathcal{F}, r)$ is

$$dist\big(P, D(\mathcal{F}, r)\big) = \min_{Q \in D(\mathcal{F},r)} \big\|H_r(P) - H_r(Q)\big\|_1 = \max\Big(0, \max_{w\colon \mathcal{B}(r) \to [0,1]} \tilde{w}(P) - m(\mathcal{F}, w)\Big). \tag{2.27}$$

The following theorem expresses that if a graph G is distinguishable with high probability from a nice family $\mathcal{F}$, then it is distinguishable based on the constant-radius neighborhood of only one random vertex, as well.

In the following theorem, we will use a function $g : \mathbb{N} \times [0,1]^2 \to \mathbb{N}$ which could be calculated from the proofs and improved by more accurate calculations, but we do not go into this direction here.

Theorem 2.10. *Assume that $H_r(G_0) \notin D(\mathcal{F}, r)$ holds for a graph $G_0 \in \mathcal{G}$ and a nice family $\mathcal{F}$ of graphs; namely, there exists a $w_0\colon \mathcal{B}(r) \to [0,1]$ satisfying*

$$\tilde{w}_0\big(H_r(G_0)\big) - m(\mathcal{F}, w_0) \ge \delta > 0. \tag{2.28}$$

Then for all $\varepsilon > 0$, with $r' = g(r, \varepsilon, \delta)$, there exists a subset $M \subset \mathcal{B}(r')$ and an induced subgraph G_1 of G_0 such that the following holds.

$$\begin{aligned} &\mathcal{P}\big(H_{r'}(G_1) \in M\big) > 1 - \varepsilon. \\ \forall G \in \mathcal{F}\colon \quad &\mathcal{P}\big(H_{r'}(G) \in M\big) < \varepsilon \end{aligned}$$

Lovász [45] asked to find, for every radius $r \in \mathbb{N}$, and error bound $\varepsilon > 0$, an explicit $n \in \mathbb{N}$ such that the r-neighborhood distribution of each graph can be ε-approximated by a graph of size at most n. Formally,

$$\forall G \in \mathcal{G}\colon \exists G' \in \mathcal{G}\colon \big|V(G')\big| \le n, \quad \big\|H_r(G), H_r(G')\big\|_1 < \varepsilon. \tag{2.29}$$

Alon gave simple proof for the existence of such a function, but the proof does not provide any explicit bound. It is still open whether, say, a recursive function exists, and also how to compute the graph G' from G. See Lovász [45] for details.

Instead of $\big|V(G')\big| \le n$, we only require that *each component* of G' has size at most n. This version is very close to the original question, because the r-neighborhood distribution of such a graph is a convex combination of the r-neighborhood distributions of the components, and each convex combination can be approximated by a graph with a bounded number of small components.

Let $n = n(r,\varepsilon)$ denote the smallest value of n satisfying the modified conditions. The following corollary shows that if there exist an arbitrary large error bound $\lambda < 1$ such that for all r, we can find an explicit upper bound on $n(r,\lambda)$, then it provides explicit upper bounds on $n(r,\varepsilon)$ for all $r \in \mathbb{N}$ and $\varepsilon > 0$, as well.

Corollary 2.11. *For all $r \in \mathbb{N}$ and $0 < \delta$ and $\lambda < 1$,*

$$n(r,\delta) \le n\big(g(r, \frac{1-\lambda}{2}, \delta), \lambda\big).$$

Proof. Let $\mathcal{F}$ be the nice family of all graphs with sizes of components at most $n(r,\delta)-1$. By the definition of $n(r,\delta)$, there exists a graph $G_0 \in \mathcal{G}$ with $dist\big(H_r(G_0), \mathcal{F}\big) \ge \delta$. Therefore, by (2.27), there exists a $w_0 \colon \mathcal{B}(r) \to [0,1]$ satisfying $\tilde{w}_0\big(H_r(G_0)\big) - m(\mathcal{F}, w_0) \ge \delta$. Let us use Theorem 2.10 with these $\mathcal{F}$, r, δ, w_0 and $\varepsilon = \frac{1-\lambda}{2}$. Let us define w_1 the characteristic function of M, namely, $w_1 : \mathcal{B}\big(g(r,\varepsilon,\delta)\big) \to \{0,1\}$ that $w_1(b) = 1$ if and only if $b \in M$. Now

$$\exists G \in \mathcal{G} \colon \forall G_1 \in \mathcal{F} \colon \qquad \tilde{w}_1\big(H_r(G)\big) - \tilde{w}_1\big(H_r(G')\big) > (1-\varepsilon) - \varepsilon = \lambda.$$

Consequently, $n(r,\delta)-1$ does not satisfy the modified conditions of (2.29) for radius $g(r, \frac{1-\lambda}{2}, \delta)$ and error bound λ. Therefore $n(r,\delta) \le n\big(g(r, \frac{1-\lambda}{2}, \delta), \lambda\big)$. □

Proof of Theorem 2.10. If a graph $G \in \mathcal{G}$ and a function $w \colon \mathcal{B}(r) \to [0,1]$ satisfies that

$$\tilde{w}\big(H_r(G)\big) = \sup_{G' \subseteq G} \tilde{w}\big(H_r(G')\big), \tag{2.30}$$

then we say that G is **supremal** for w.

For a graph $G \in \mathcal{G}$, radius $r \in \mathbb{N}$, weighting $w \colon \mathcal{B}(r) \to [0,1]$ and $\alpha > 0$, let $A = A(G,r,w,\alpha) = \big(G = (S,R,T,\mathbf{E}), c\big)$ denote the following (**a**veraging) network. S, R, T are three copies of $V(G)$, and for each $v \in V(G)$, we denote its copies by v_S, v and v_T, respectively. We identify R with $V(G)$. Let $\mathbf{E} = \mathbf{E}(G) \cup \big\{(v_S, v), (v, v_S), (v, v_T), (v_T, v) \;\big|\; v \in V(G)\big\}$. For each $e \in \mathbf{E}(G)$, let $c(e) = d^r$, and for each vertex $v \in V(G)$, let $c(v_S, v) = w\big(h_r(G,v)\big)$ and $c(v, v_T) = \alpha$, and $c(v, v_S) = c(v_T, v) = 0$.

Lemma 2.12. *For any graph $G \in \mathcal{G}$ and function $w \colon \mathcal{B}(r) \to [0,1]$ and $\alpha > 0$, the size of the maximum flow $\big\|f^*\big\|$ in $A(G,r,w,\alpha)$ satisfies*

$$\min\Big(\alpha,\ \inf_{G' \subseteq G} \tilde{w}\big(H_r(G')\big)\Big)n \le \big\|f^*\big\| \le^{(*)} \min\Big(\alpha,\ \tilde{w}\big(H_r(G)\big)\Big)n \tag{2.31}$$

with equation at (*) if G is supremal for w.

Proof. Notice that for any graph $G \in \mathcal{G}$,

$$\tilde{w}\big(H_r(G)\big) = \frac{1}{|V(G)|} \sum_{v \in V(G)} w\big(h_r(G,v)\big). \tag{2.32}$$

Consider the following two cuts. $\|S \cup R\| = \alpha n$ and $\|S\| = \tilde{w}\big(H_r(G)\big)n$, which proves the upper bound of (2.31).

On the other hand, consider an arbitrary cut X of A. Let $R^- = R \cap X$ and $R^+ = R \backslash X$. Let G^- and G^+ denote the subgraphs of G induced by R^- and R^+, respectively. Let δ_X denote the number of edges between R^- and R^+. For each edge between R^- and R^+, its r-neighborhood contains at most d^r vertices. Therefore there exist at most $d^r\delta$ vertices $v \in R^-$ for which $h_r(G^-, v) \neq h_r(G, v)$, and at most $d^r\delta$ vertices $v \in R^+$ for which $h_r(G^+, v) \neq h_r(G, v)$. These imply the following inequalities.

$$\sum_{v \in R^-} \Big(w\big(h_r(G, w)\big) - w\big(h_r(G^-, w)\big) \Big) \leq d^r \delta_X. \tag{2.33}$$

$$\sum_{v \in R^+} \Big(w\big(h_r(G^+, w)\big) - w\big(h_r(G, w)\big) \Big) \leq d^r \delta_X. \tag{2.34}$$

Now we prove the lower bound of (2.31).

$$\|X\| \overset{(2.5)}{=} |R^-|\alpha + \sum_{v \in R^+} w\big(h_r(G, v)\big) + d^r\delta_X \overset{(2.34)}{\geq} |R^-|\alpha + \sum_{v \in R^+} w\big(h_r(G^+, v)\big)$$

$$\overset{(2.32)}{=} |R^-|\alpha + |R^+|\tilde{w}\big(H_r(G^+)\big) \geq \Big(|R^-| + |R^+|\Big) \min\Big(\alpha,\ \tilde{w}\big(H_r(G^+)\big)\Big)$$

$$= \min\Big(\alpha,\ \tilde{w}\big(H_r(G^+)\big)\Big)n \geq \min\Big(\alpha,\ \inf_{G' \subseteq G} \tilde{w}\big(H_r(G')\big)\Big)n.$$

Finally, we show that if G is supremal for w, then the upper bound of (2.31) is also a lower bound.

$$\|X\| \overset{(2.5)}{=} |R^-|\alpha + \sum_{v \in R^+} w\big(h_r(G, v)\big) + d^r\delta_X = |R^-|\alpha + \sum_{v \in R} w\big(h_r(G, v)\big) - \sum_{v \in R^-} w\big(h_r(G, v)\big)$$

$$+ d^r\delta_X \overset{(2.33)}{\geq} |R^-|\alpha + \sum_{v \in R} w\big(h_r(G, v)\big) - \sum_{v \in R^-} w\big(h_r(G^-, v)\big) \overset{(2.32)}{=} |R^-|\alpha + |R|\tilde{w}\big(H_r(G)\big)$$

$$- |R^-|\tilde{w}\big(H_r(G^-)\big) = |R^-|\alpha + |R^+|\tilde{w}\big(H_r(G)\big) + |R^-|\Big(\tilde{w}\big(H_r(G)\big) - \tilde{w}\big(H_r(G^-)\big)\Big) \overset{(2.30)}{\geq}$$

$$|R^-|\alpha + |R^+|\tilde{w}\big(H_r(G)\big) \geq \Big(|R^-| + |R^+|\Big) \min\Big(\alpha,\ \tilde{w}\big(H_r(G)\big)\Big) \geq \min\Big(\alpha,\ \tilde{w}\big(H_r(G)\big)\Big)n. \quad \square$$

For a radius r, weighting $w\colon \mathcal{B}(r) \to [0, 1]$, $\varepsilon > 0$ and $\alpha > 0$, we define the following operator $W_\varepsilon(r, w, \alpha)$. Its value will be a new weighting $w'\colon \mathcal{B}(r') \to [0, 1]$, where $r' = g_1(r, \varepsilon)$ is r plus the radius used in Theorem 2.1 with error bound ε/d^r.

Consider an arbitrary $B \in \mathcal{B}(r')$ with root v. Let $A_0 = A(B, r, w, \alpha)$. Consider the flow f generated by the local flow algorithm in Theorem 2.1 on A_0 with error bound ε. (We divide the capacities by d^r, calculate the flow with error bound ε/d^r and then multiply the flow by d^r.) Then we define

$$w'(B) = \alpha - f(v, v_T). \tag{2.35}$$

Notice that

$$w'(B) = \alpha - f(v, v_T) \in \alpha - [0, \alpha] = [0, \alpha]. \tag{2.36}$$

Lemma 2.13. *For each graph G, radius r, weighting $w\colon \mathcal{B}(r) \to [0,1]$, $\varepsilon > 0$, $\alpha > 0$, $r' = g_1(r,\varepsilon)$ and $w' = W_\varepsilon(r, w, \alpha)$,*

$$\max\Big(0,\ \alpha - \tilde{w}\big(H_r(G)\big)\Big) \le \tilde{w}'\big(H_{r'}(G)\big) \le \max\Big(0,\ \alpha - \inf_{G'\subseteq G} \tilde{w}\big(H_r(G')\big)\Big) + \varepsilon. \tag{2.37}$$

Furthermore, if G is supremal for w, then the lower bound is ε-tight, namely

$$\tilde{w}'\big(H_{r'}(G)\big) \le \max\Big(0,\ \alpha - \tilde{w}\big(H_r(G)\big)\Big) + \varepsilon. \tag{2.38}$$

Proof. Let $A_1 = A(G, r, w, \alpha)$, and consider the flow f generated by the local flow algorithm in Theorem 2.1 on A_1 with error bound ε. For each $v \in V(G)$, $f(v, v_T)$ depends only on $h_{r'}(G, v)$, therefore (2.35) shows that

$$w'\big(h_{r'}(G,v)\big) = \alpha - f(v, v_T). \tag{2.39}$$

Hence we get that

$$\begin{aligned}
\tilde{w}'\big(H_{r'}(G)\big) &\overset{(2.32)}{=} \frac{1}{n}\sum_{v\in V(G)} w'\big(h_r(G,v)\big) \overset{(2.39)}{=} \frac{1}{n}\sum_{v\in V(G)} \big(\alpha - f(v,v_T)\big) \\
&= \alpha - \frac{1}{n}\sum_{v\in V(G)} f(v,v_T) = \alpha - \frac{1}{n}\big\|f\big\| \overset{(2.7)}{\in} \alpha - \frac{1}{n}\Big[\big\|f^*\big\| - \varepsilon n, \big\|f^*\big\|\Big] \\
&\overset{(2.31)}{\subseteq} \alpha - \left[\min\Big(\alpha,\ \inf_{G'\subseteq G} \tilde{w}\big(H_r(G')\big)\Big) - \varepsilon,\ \min\Big(\alpha,\ \tilde{w}\big(H_r(G)\big)\Big)\right] \\
&= \left[\max\Big(0,\ \alpha - \tilde{w}\big(H_r(G)\big)\Big),\ \max\Big(0,\ \alpha - \inf_{G'\subseteq G} \tilde{w}\big(H_r(G')\big)\Big) + \varepsilon\right]
\end{aligned}$$

□

For

$$\varepsilon_1 = \delta\varepsilon/2, \tag{2.40}$$

let $w_1 = W_{\varepsilon_1}\big(r, w_0, \tilde{w}_0(H_r(G_0))\big)$ with radius r_1. Then, let $w_2 = W_{\varepsilon_1}\big(r_1, w_1, \delta\big)$, and let r' be the radius it uses. Let us define

$$M = \big\{B \in \mathcal{B}_{r'} : w_2(B) > \delta/2\big\}. \tag{2.41}$$

Let G_1 be the induced subgraph of G_0 with the lowest $\tilde{w}_1\big(H_{r_1}(G_1)\big)$. We show that this satisfies the requirements.

Let G_2 be the induced subgraph of G_0 with the highest $\tilde{w}_0\big(H_r(G_2)\big)$. These also imply G_1 is infimal for w_1, and G_2 is supremal for w_0, namely

$$\forall G' \subseteq G_0 : \qquad \tilde{w}_1\big(H_{r_1}(G')\big) \ge \tilde{w}_1\big(H_{r_1}(G_1)\big) = \inf_{G\subseteq G_1} \tilde{w}_1\big(H_{r_1}(G)\big), \tag{2.42}$$

$$\forall G' \subseteq G_0 : \qquad \tilde{w}_0\big(H_r(G')\big) \le \tilde{w}_0\big(H_r(G_2)\big) = \sup_{G\subseteq G_2} \tilde{w}_0\big(H_r(G)\big). \tag{2.43}$$

$$\begin{aligned}
\tilde{w}_2\big(H_{r'}(G_1)\big) &\overset{(2.37)}{\ge} \delta - \tilde{w}_1\big(H_{r_1}(G_1)\big) \overset{(2.42)}{\ge} \delta - \tilde{w}_1\big(H_{r_1}(G_2)\big) \\
&\overset{(2.38)}{\ge} \delta - \left(\max\Big(0,\ \tilde{w}_0\big(H_r(G_0)\big) - \tilde{w}_0\big(H_r(G_2)\big)\Big) + \varepsilon_1\right) \overset{(2.43)}{\ge} \delta - \varepsilon_1.
\end{aligned} \tag{2.44}$$

$$\mathcal{P}\big(H_{r'}(G_1) \in M\big) \overset{(2.41)}{=} \mathcal{P}\Big(w_2\big(H_{r'}(G_1)\big) > \delta/2\Big) = \mathcal{P}\Big(\delta - w_2\big(H_{r'}(G_1)\big) \le \delta/2\Big) \tag{2.45}$$

(2.36) shows that $\forall B \in \mathcal{B}_{r'}\colon w_2(B) \leq \delta$, therefore using Markov's inequality for $\delta - w_2\big(H_{r'}(G_1)\big) > 0$,

$$(2.45) \geq 1 - \frac{\mathsf{E}\big(\delta - w_2(H_{r'}(G_1))\big)}{\delta/2} = 1 - \frac{\delta - \mathsf{E}\big(w_2(H_{r'}(G_1))\big)}{\delta/2}$$

$$\overset{(2.25)}{=} 1 - \frac{\delta - \tilde{w}_2\big(H_{r'}(G_1)\big)}{\delta/2} \overset{(2.44)}{\geq} 1 - \frac{\delta - (\delta - \varepsilon_1)}{\delta/2} = 1 - \frac{2\varepsilon_1}{\delta} \overset{(2.40)}{=} 1 - \varepsilon.$$

On the other hand, for all $G \in \mathcal{F}$,

$$\tilde{w}_1\big(H_{r_1}(G)\big) \overset{(2.37)}{\geq} \tilde{w}_0\big(H_r(G_0)\big) - \tilde{w}_0\big(H_r(G)\big) \overset{(2.26)}{\geq} \tilde{w}_0\big(H_r(G_0)\big) - m(\mathcal{F}, w_0) \overset{(2.28)}{\geq} \delta.$$

Therefore, as $\mathcal{F}$ is closed under taking induced subgraphs, for all $G \in \mathcal{F}$,

$$\inf_{G \subseteq G_1} \tilde{w}_1\big(H_{r_1}(G)\big) \geq \delta. \tag{2.46}$$

Therefore for all $G \in \mathcal{F}$,

$$\tilde{w}_2\big(H_{r'}(G)\big) \overset{(2.37)}{\leq} \max\Big(0,\ \delta - \inf_{G \subseteq G_1} \tilde{w}_1\big(H_{r_1}(G)\big)\Big) + \varepsilon_1 \overset{(2.46)}{=} \varepsilon_1. \tag{2.47}$$

Using Markov's inequality for $w_2\big(H_{r'}(G)\big) > 0$,

$$\mathcal{P}\big(H_{r'}(G) \in M\big) \overset{(2.45)}{=} \mathcal{P}\Big(w_2\big(H_{r'}(G)\big) > \delta/2\Big) \leq \frac{\mathsf{E}\big(w_2(H_{r'}(G))\big)}{\delta/2} \overset{(2.25)}{=} \frac{\tilde{w}_2\big(H_{r'}(G)\big)}{\delta/2} \overset{(2.47)}{\leq} \frac{2\varepsilon_1}{\delta} \overset{(2.40)}{=} \varepsilon.$$

□

2.3.1 Connection with the Aldous-Lyons conjecture

There is a large theory about the Aldous-Lyons conjecture which we do not present here, but we try to give a bit foggy description about it just to show the applicability of our algorithm. For those who are interested in the Aldous-Lyons conjecture, we suggest reading [1] or [45].

Consider a probability distribution U of rooted graphs, including infinite graphs, with degrees bounded by d. Select a connected rooted graph from U, and then select a uniform random edge e from the root. We consider e as oriented away from the root. This way we get a probability distribution σ with an oriented "root edge". Let $R(\sigma)$ denote the distribution obtained by reversing the orientation of the root of a random element of σ. We say that U is a **unimodular** random graph if $\sigma = R(\sigma)$.

Let $\mathcal{U}$ denote the set of unimodular random graphs. Let $H_r(U)$ denote the distribution of the r-neighborhoods of the root of $U \in \mathcal{U}$. Let $D(\mathcal{U}, r) = cl\big\{H_r(U)\big|U \in \mathcal{U}\big\}$. It can be easily shown that every graph G with a uniform random root provides a $U \in \mathcal{U}$ with $H_r(G) = H_r(U)$, therefore $D(\mathcal{G}, r) \subseteq D(\mathcal{U}, r)$.

The Aldous–Lyons conjecture say that every unimodular distribution on rooted connected graphs with bounded degree is the "limit" of a bounded degree graph sequence. More precisely,

Conjecture 2.14 (Aldous–Lyons)**.**

$$\forall r \in \mathbb{N}\colon\ D(\mathcal{G}, r) = D(\mathcal{U}, r)(r).$$

Definition 1. *We say that the Aldous–Lyons Conjecture is* ***completely false*** *if*

Statement/Conjecture 2.15. *If the Aldous–Lyons Conjecture is false, then there exists a unimodular random graph U that can be distinguished with high probability from any graph, based on the constant-radius neighborhood of only one random vertex. With other words, for all $\varepsilon > 0$, there exists an $r \in \mathbb{N}$, a subset $M \subset \mathcal{B}(r)$ and a unimodular random graph $U \in \mathcal{U}$ that for all $G \in \mathcal{G}$, the r-neighborhood of a random vertex of G is in M with probability at most ε, but the r-neighborhood of the root of U is in M with probability at least $1 - \varepsilon$.*

Justification. If $D(\mathcal{G}, r) \subsetneq D(\mathcal{U}, r)$, then there exists a function $w\colon \mathcal{B}(r) \to [0, 1]$ so that

$$\sup_{G \in \mathcal{G}} \tilde{w}\big(H_r(G)\big) < \sup_{U \in \mathcal{U}} \tilde{w}\big(H_r(U)\big).$$

We only need to prove Theorem 2.10 with $G_0 \in U$ instead of $G_0 \in \mathcal{G}$. The author believes that an analogous proof works here, but many technical details should be changed, and the complete proof would be the topic of another paper. For example, we do not have Maximum Flow–Minimum Cut Theorem for unimodular random graphs, but we can deduce the Supremum Flow–Infimum Cut Theorem from the local algorithms presented here. □

Chapter 3

Local algorithms

In this chapter, we show that preprocession is useless for local algorithms. More precisely, if there exists a local algorithm using preprocession, then there exists another local algorithm with the same radius, in which the only "preprocession" is a random variable with a continuous distribution, and this provides an output with at most the same error from the optimum, in expectation.

3.1 Model and results

Graph means finite graph with degrees bounded by a fixed constant, allowing loops and parallel edges. Whenever we take a function depending on a graph, we mean that it depends on the isomorphism type of the graph. In other words, graphs are considered as unlabelled. The r-neighborhood of a vertex x of a graph G, denoted by $B_r(x)$ or $B_r(G,x)$, means the rooted subgraph of G induced by all nodes at distance at most r from x, and rooted at x. For a family $\mathcal{F}$ of graphs, denote the family of rooted r-neighborhoods by $\mathcal{F}_r = \big\{B_r(G,x)\big|G \in \mathcal{F};\ x \in V(G)\big\}$. For any sequence of graphs G_i, let $\bigcup G_i$ denote their disjoint union, that is, $V(\bigcup G_i) = \big\{(x,i)\big|x \in G_i\big\}$ and $E(\bigcup G_i) = \Big\{\big((x,i),(y,i)\big)\Big|(x,y) \in \mathsf{E}(G_i)\Big\}$. A 5-tuple $(\mathcal{F}, C, \delta, \mathcal{A}, v)$ is called a **local choice problem**, where

- $\mathcal{F}$ is a union-closed family of graphs, that is, $G, H \in \mathcal{F} \Rightarrow G \cup H \in \mathcal{F}$;
- C is an arbitrary set (the image range of choices);
- δ is a positive integer (the radius);
- $\mathcal{A}$ is a set of pairs (H, c) where $H \in \mathcal{F}_\delta$ and c is a function $V(H) \to C$ (the set of allowable choices on neighborhoods);
- v is a function $C \to (-\infty, M]$ (the valuation of a choice).

Let **choice** mean a function $c\colon V(G) \to C$. Given a graph G, we call a choice c **allowed** if $\forall x \in V(G)\colon \big(B_\delta(G,x), c|_{V(B_\delta(G,x))}\big) \in \mathcal{A}$. We denote the set of all allowed choices by $\mathcal{A}(G)$. The **value of a choice** is

$$\bar{v}(G,c) = \frac{1}{|V(G)|} \sum_{x \in V(G)} v\big(c(x)\big), \tag{3.1}$$

and the value of a graph is

$$v^*(G) = \sup_{c \in \mathcal{A}(G)} \bar{v}(G,c). \tag{3.2}$$

Given a local choice problem, our aim is to find an allowed c for every graph G with $\bar{v}(G,c)$ close to $v^*(G)$.

For example, one way to describe the maximum matching problem in this language is the following. $\mathcal{F}$ is the family of all graphs; $C = [0,1] \cup \{\emptyset\}$; $\delta = 1$; $(H,c) \in \mathcal{A}$ iff for the root x of H; $c(x) = \emptyset$ or there exists exactly 1 neighbor y of x with $c(x) = c(y)$; Finally, $v(col)$ is 0 if $col = \emptyset$ and $v(col) = \frac{1}{2}$ otherwise. Then the allowed choices describe the matchings: $c(x) = \emptyset$ if x is unmatched, otherwise x is matched with the neighboring vertex y with $c(x) = c(y)$. Now, $\bar{v}(G,c)$ describes the size of this matching, normalized by $\big|V(G)\big|$.

For the first look, it would be more natural and general to evaluate a given choice based on the choices in the δ-neighborhood of the node, not just on the choice at the node. In other words, we could define $v\colon \mathcal{A} \to (-\infty, M]$ and $\bar{v}(G,c) = \frac{1}{|V(G)|} \sum_{x \in V(G)} v\big(B_\delta(G,x), c|_{B_\delta(G,x)}\big)$. In fact, this definition would not be more general than the original version. Roughly, because we can define the coloring so as to include the value of the coloring at the point. More formally, let $(\mathcal{F}, C, \delta, \mathcal{A}, v)$ be an extended local choice problem, where we use this more general v. Let $C' = C \times \mathbb{R}$ and $\mathcal{A}' = \Big\{ \big(H, (c_1,c_2)\big) \Big| (H,c_1) \in \mathcal{A};\ c_2\big(root(H)\big) = v(H,c_1) \Big\}$ and $v'\big(H,(c_1,c_2)\big) = c_2$. Then the local choice problem $\big(\mathcal{F}, C', \delta, \mathcal{A}', v'\big)$ is equivalent in an appropriate sense to the extended local choice problem $(\mathcal{F}, C, \delta, \mathcal{A}, v)$. The details are left to the Reader.

Now we define different versions of local algorithms for finding such an allowed choice c. We assign independent identically distributed random variables to the vertices with a fixed distribution D. We denote this random assignment by $\omega\colon V(G) \to \Omega$. The most important case is when D is a continuous distribution, say uniform on $[0,1]$, but it can be a constant number of random bits, or an arbitrary distribution. We take one more independent public random variable g with an arbitrary distribution.

[**Preprocessed**] [**Mixed**] [**Random**] **Local Algorithm** ([**P**][**M**][**R**]**LA**) means a function that assigns a choice c to each graph G, in the following way. There is a fix radius r that for each G and $x \in V(G)$, the algorithm sets $c(x)$ depending on $B_r(x)$

- and the isomorphism type of the graph G if Preprocessed,
- and on the global random seed g if Mixed,
- and on the local random seeds in the r-neighborhood $\omega|_{V(B_r(x))}$ if Random.

Now we have $2 \times 2 \times 2$ different versions of local algorithms. We say that an algorithm is correct if it always[1] produces allowed choices. The main result of this paper is that MRLA and PMRLA are equally strong, in the following sense. We say that a local choice problem is approximable by a **t**ype of **a**lgorithm TA if, for all $\varepsilon > 0$, there exists a correct TA f that $\forall G \in \mathcal{F}\colon \mathsf{E}\big(\bar{v}(G,f)\big) \geq v^*(G) - \varepsilon$.

Theorem 3.1. *If a local choice problem is approximable by PMRLA, then this is approximable by MRLA, as well.*

The most general form of this result is the following. Let $l[G,\omega,g]$ denote the choice on $G \in \mathcal{F}$ computed by the MRLA l and the vector of the local random seeds ω and the global random seed g.

Theorem 3.2. *Let $b\colon \mathbb{R} \to \mathbb{R}$ be a monotone increasing concave function and $\varepsilon > 0$. If there exists a correct PMRLA l_1 such that, for each graph $G \in \mathcal{F}$, $\mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l_1[G,\omega,g]\big)\Big) \geq b\big(v^*(G)\big)$,*

[1] We could use "with probability 1" instead of "always" with essentially the same proofs.

then there exists a correct MRLA l_2 using the same radius r and the same distribution of ω so that $\mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l_2[G,\omega,g]\big)\Big) > b\big(v^*(G)\big) - \varepsilon$.

Remember that the distribution D of ω is set arbitrarily but fixed. Therefore, [P][M] random local algorithms can have different strength depending on D, but we always use the same D. For example, applying the two theorems when D is a constant distribution, we get the same results for the equistrength of PMLA and MLA.

At the end of the paper, we will show that preprocessing and mixing are just equally strong. We will also show that there are problems approximable by MLA but not by RLA, and vice versa.

3.2 Proof of Theorems 3.1 and 3.2

Without loss of generality, we assume that $r \geq \delta$. Denote by $s_r(G)$ the exact distribution of $B_r(G,x)$ for a uniform random vertex $x \in V(G)$, in other words,

$$\forall H \in \mathcal{F}_r\colon \qquad s_r(G)(H) = \Big|\big\{x \big| B_r(G,x) \cong H\big\}\Big| / \big|V(G)\big|.$$

Lemma 3.3. *Given a MRLA l using radius r, $\mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l[G,\omega,g]\big)\Big)$ is a linear function of $s_r(G)$.*

Proof. We average on a random variable upper bounded by M, therefore, its expected value exists.

$$\mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l[G,\omega,g]\big)\Big) \overset{(3.1)}{=} \mathsf{E}_{\omega,g}\Big(\frac{1}{|V(G)|}\sum_{x\in V(G)} v\big(l[G,\omega,g](x)\big)\Big)$$

$$= \frac{1}{|V(G)|}\sum_{x\in V(G)} \mathsf{E}_{\omega,g}\Big(v\big(l[G,\omega,g](x)\big)\Big). \tag{3.3}$$

Notice that $v\big(l[G,\omega,g](x)\big)$ depends only on $B_r(G,x)$, ω and g. Therefore, $\mathsf{E}_{\omega,g}\Big(v\big(l[G,\omega,g](x)\big)\Big)$ depends only on $B_r(G,x)$. Let

$$p_l\big(B_r(G,x)\big) = \mathsf{E}_{\omega,g}\Big(v\big(l[G,\omega,g](x)\big)\Big). \tag{3.4}$$

Continuing the calculations,

$$(3.3) \overset{(3.4)}{=} \frac{1}{|V(G)|}\sum_{x\in V(G)} p_l\big(B_r(G,x)\big) = \sum_{H\in\mathcal{F}_r} s_r(G)(H)\cdot p_l(H). \qquad \square$$

The following technical lemma is not about the essence of the proof, but required.

Lemma 3.4. *For each local choice problem and radius r, there exists a graph $T_r \in \mathcal{F}$ so that for all $G \in \mathcal{F}$, the following holds. If a MRLA l with radius r produces an allowed choice on $T_r \cup G$, then l is correct.*

Proof. For each $H \in \mathcal{F}_r$, let us choose a graph $a(H)$ so that $\exists x \in V\big(a(H)\big)\colon B_{\delta+r}\big(a(H),x\big) \cong H$. We show that $T_r = \bigcup_{H\in\mathcal{F}_r} a(H)$ satisfies the requirement.

Suppose that a MRLA l is not correct. This means that there exists $G \in \mathcal{F}$ and $x \in V(G)$ and $\omega\colon V(G) \to supp(\Omega)$ and g such that $\Big(B_\delta(x), l[G,\omega,g]\big|_{V(B_\delta(x))}\Big) \notin \mathcal{A}$. For each $y \in V\big(B_\delta(x)\big)$, $l[G,\omega,g](y)$ depends only on $B_{\delta+r}(y)$ and $\omega|_{V(B_{\delta+r}(y))}$ and g. $V\big(B_r(y)\big) \subseteq V\big(B_{\delta+r}(x)\big)$, so $\Big(B_\delta(x), l[G,\omega,g]\big|_{V(B_\delta(x))}\Big)$ depends only on $B_{\delta+r}(x)$ and $\omega|_{V(B_{\delta+r}(x))}$ and g. Thus, if we take the component $a\big(B_{\delta+r}(x)\big)$ of T_r and the same ω on $B_{\delta+r}(x')$ (x' is the vertex in T_r corresponding to x in $B_{\delta+r}(x)$) and the same g, then it produces the same pair $\Big(B_\delta(T_r,x'), l[T_r,\omega,g]\big|_{V(B_\delta(T_r,x'))}\Big) \cong \Big(B_\delta(x), l[G,\omega,g]\big|_{V(B_\delta(G,x))}\Big) \notin \mathcal{A}$.

So the choice produced by l on $T_r \cup G$ is not allowed. □

Based on that

$$v^*(G \cup H) = \frac{|V(G)|v^*(G) + |V(H)|v^*(H)}{|V(G)| + |V(H)|} \text{ and } s_r(G \cup H) = \frac{|V(G)|s_r(G) + |V(H)|s_r(H)}{|V(G)| + |V(H)|},$$

we show the following lemma.

Lemma 3.5. *The set $S_r = cl\big(\{s_r(G)\big|G \in \mathcal{F}\}\big)$ is convex, and the function $m_r\colon S_r \to \mathbb{R}$ defined by*

$$m_r(q) = \limsup_{G_n \in \mathcal{F}, s_r(G_n) \to q} v^*(G_n) \tag{3.5}$$

is concave.

Proof. For an integer k and a graph G, let $k \times G$ denote $\bigcup_{i=1}^{k} G_i$, where each $G_i \cong G$. For choices $c_i\colon V(G_i) \to C$, let $\bigcup_{i=1}^{k} c_i\colon \bigcup_{i=1}^{k} V(G_i) \to C$ denote the function $(\bigcup_{i=1}^{k} c_i)\big((x,j)\big) = c_j(x)$. For a choice $c\colon V(G) \to C$, let $k \times c = \bigcup_{i=1}^{k} c_i$, where each $c_i\colon G_i \to C$ is a copy of $c\colon G \to C$.

Let $q_0, q_1 \in S_r$, and for all $\lambda \in [0,1]$, $q_\lambda = (1-\lambda)\cdot q_0 + \lambda \cdot q_1$.

$$(1-\lambda)\cdot m(q_0) + \lambda \cdot m(q_1) \overset{(3.5)}{=} (1-\lambda)\cdot \limsup_{s_r(G_n)\to q_0} v^*(G_n) + \lambda \cdot \limsup_{s_r(G_n)\to q_1} v^*(G_n)$$

$$\overset{(3.2)}{=} (1-\lambda)\cdot \limsup_{s_r(G_n)\to q_0} \sup_{c \in \mathcal{A}(G_n)} \bar{v}(G_n,c) + \lambda \cdot \limsup_{s_r(G_n)\to q_1} \sup_{c \in \mathcal{A}(G_n)} \bar{v}(G_n,c)$$

$$= \limsup \Big\{ (1-\lambda)\cdot \bar{v}(G_n^{(0)}, c_n^{(0)}) + \lambda \cdot \bar{v}(G_n^{(1)}, c_n^{(1)}) \Big| \forall i \in \{0,1\}\colon \big(s_r(G_n^{(i)}) \to q_i;\ c_n^{(i)} \in \mathcal{A}(G_n^{(i)})\big) \Big\}$$

$$= \limsup \Big\{ \frac{b_n - a_n}{b_n} \cdot \bar{v}(G_n^{(0)}, c_n^{(0)}) + \frac{a_n}{b_n} \cdot \bar{v}(G_n^{(1)}, c_n^{(1)})$$

$$\Big| a_n, b_n \in \mathbb{N};\ \frac{a_n}{b_n} \to \lambda;\ \forall i \in \{0,1\}\colon \big(s_r(G_n^{(i)}) \to q_i;\ c_n^{(i)} \in \mathcal{A}(G_n^{(i)})\big) \Big\}$$

$$= \limsup \Big\{ \bar{v}\Big((b_n - a_n)\big|V(G^{(1)})\big| \times G_n^{(0)} \bigcup a_n\big|V(G^{(0)})\big| \times G_n^{(1)}, (b_n - a_n)\big|V(G^{(1)})\big| \times c_n^{(0)}$$

$$+ a_n\big|V(G^{(0)})\big| \times c_n^{(1)}\Big) \Big| a_n, b_n \in \mathbb{N};\ \frac{a_n}{b_n} \to \lambda;\ \forall i \in \{0,1\}\colon \big(s_r(G_n^{(i)}) \to q_i;\ c_n^{(i)} \in \mathcal{A}(G_n^{(i)})\big) \Big\}$$

It is easy to check that $s_r\Big((b_n-a_n)\big|V(G^{(1)})\big|\times G_n^{(0)}+a_n\big|V(G^{(0)})\big|\times G_n^{(1)})\Big)=q_{a_n/b_n}\to q_\lambda$. This implies the convexity of S_r, and continuing the calculations,

$$\leq \limsup\big\{\bar v(G_n,c_n)\big|s_r(G_n)\to q_\lambda;\ c_n\in\mathcal{A}(G_n)\big\}\overset{(3.2)}{=}\limsup_{s_r(G_n)\to q_\lambda} v^*(G_n)\overset{(3.5)}{=}m(q_\lambda),$$

which means the concavity of m. □

Lemma 3.6. *Given a compact convex set $X\subset\mathbb{R}^n$, and two convex functions $f_0,f_1\colon X\to\mathbb{R}$ that for each $x\in X\colon f_0(x)>0$ or $f_1(x)>0$. Then there exists a convex combination of the functions, which is positive on each point in X. Formally,*

$$\exists\lambda\in[0,1]\colon\forall x\in X\colon\qquad f_\lambda(x)=\big((1-\lambda)\cdot f_0+\lambda\cdot f_1\big)(x)>0.$$

Proof. Let $f_\lambda^-=\big\{x\in X\big|f_\lambda(x)\leq 0\big\}$. Each f_λ^- is convex and compact, and f_0^- and f_1^- are disjoint. If $f_0(x)>0$ and $f_1(x)>0$, then $f_\lambda(x)>0$ as well, so $f_\lambda^-\subseteq f_0^-\cup f_1^-$. These together ($f_0^-$ and f_1^- are compact and disjoint, f_λ^- is convex, $f_\lambda^-\subseteq f_0^-\cup f_1^-$) imply that $f_\lambda^-\subseteq f_0^-$ or $f_\lambda^-\subseteq f_1^-$.

For any compact set S, the function $\lambda\to\min\limits_{x\in S}f_\lambda(x)$ is continuous, so $\big\{\lambda\in[0,1]\big|\min\limits_{x\in S}f_\lambda(x)\leq 0\big\}$ is closed. Therefore, the sets $A=\big\{\lambda\in[0,1]\big|f_\lambda^-\cap f_0^-\neq\emptyset\big\}$ and $B=\big\{\lambda\in[0,1]\big|f_\lambda^-\cap f_1^-\neq\emptyset\big\}$ are closed, disjoint and nonempty. Thus $A\cup B$ cannot be $[0,1]$, because $[0,1]$ is a connected topological space. Therefore, there exists a $\lambda\in[0,1]-A-B$, and this satisfies the requirements. □

Lemma 3.7. *Given a compact convex set $X\subset\mathbb{R}^n$. For each $x\in X$, there is a given convex function f_x so that $f_x(x)>0$. Then there exists a convex combination of the functions that is positive on each point in X. Formally, there exist $x_1,x_2,...\in X;\ \lambda_1,\lambda_2,...\geq 0;\ \sum\limits_i\lambda_i=1$ so that $\forall y\in X\colon\sum\limits_i\lambda_i f_{x_i}(y)>0$.*

Proof. Consider the set $\mathcal{T}$ of convex combinations of f_x -s. Each function in this set is convex. For each function $h\in\mathcal{T}$, let us call $h^+=\big\{x\in X\big|h(x)>0\big\}$ and $h^-=\big\{x\in X\big|h(x)\leq 0\big\}$ the positive and the nonpositive set of h, respectively. The positive set of each function is open, and these cover together the compact set X. This implies that there exists finitely many of these functions such that their positive sets cover X. Consider a smallest family: $h_1,h_2,...\,h_n$.

Assume that $n>1$. The nonpositive set of a function is convex and compact. Let $X'=X\cap h_3^-\cap h_4^-\cap...\,h_n^-$. This is the intersection of finitely many convex compact sets, so X' is convex and compact, as well. At each point $x\in X'$, $h_1(x)>0$ or $h_2(x)>0$, otherwise x would not be covered by any h_i^+. Therefore, Lemma 3.6 shows that there exists a convex combination h_0 of h_1 and h_2 which is positive on X'. Clearly, $h_0\in\mathcal{T}$ and $h_0^+\cup h_3^+\cup h_4^+\cup...\,h_n^+=h_0^+\cup(X-X')=X$, contradicting the assumption that n is the smallest number of functions required. Therefore, $n=1$, which means that this function is positive on X. □

Lemma 3.8. *Given a compact convex set $X\subset\mathbb{R}^n$, and a concave function $f\colon X\to\mathbb{R}$. For each $x\in X$, there is a given linear function f_x so that $f_x(x)>f(x)$. Then there exists a convex combination of f_x -s upper bounding f. Formally $\exists x_1,x_2,...\in X;\ \lambda_1,\lambda_2,...\geq 0;\ \sum\limits_i\lambda_i=1$ so that*

$$\forall y\in X\colon\qquad\sum_i\lambda_i f_{x_i}(y)>f(y).\tag{3.6}$$

Proof. The functions $f_x - f$ are convex. $f_x(x) > f(x)$ is equivalent to $(f_x - f)(x) > 0$. If $\sum_i \lambda_i = 1$, then $\sum_i \lambda_i f_{x_i}(y) > f(y)$ is equivalent to $\sum_i \lambda_i \big(f_{x_i} - f\big)(y) > 0$. So we can use Lemma 3.7 with the functions $f_x - f$ (as f_x there) and it gives the convex combination satisfying the requirement. □

Proof of Theorem 3.2. Let $D(\mathcal{F}_r)$ denote the finite dimensional space of the probability distributions on $\mathcal{F}_r$. For a MRLA l and a $q \in D(\mathcal{F}_r)$ and using the notation in Lemma 3.3, let

$$u(l, q) = \sum_{H \in \mathcal{F}_r} q(H) \cdot p_l(H). \tag{3.7}$$

Clearly, $v \le M$, thus $p_l \le M$, thus $u \le M$.

We use the graph T_r defined in Lemma 3.4. Let $\tilde{\mathcal{F}}[r] = \{G \cup T_r | G \in \mathcal{F}\} \subseteq \mathcal{F}$. Let $\tilde{S}_r = \big\{s_r(G)\big|G \in \tilde{\mathcal{F}}[r]\big\}$. Then we have $\lim_{n\to\infty} s_r(n \times G \cup T_r) = s_r(G)$, therefore, $cl(\tilde{S}_r) = cl\big(\big\{s_r(G)\big|G \in \tilde{\mathcal{F}}[r]\big\}\big) = cl\big(\big\{s_r(G)\big|G \in \mathcal{F}\big\}\big) = S_r$.

Given G, the PMRLA is a MRLA. Lemma 3.4 implies that if $G \in \tilde{\mathcal{F}}[r]$, then this MRLA is correct not only for G but for all graphs. Therefore, given an arbitrary $G \in \tilde{\mathcal{F}}[r]$, there exists a correct MRLA $l = l_{s_r(G)}$ with

$$u\big(l, s_r(G)\big) \ge b\Big(m_r\big(s_r(G)\big)\Big). \tag{3.8}$$

Consider an arbitrary $q \in S_r$. Let $\tilde{m}_r(q) = \limsup_{G_n \in \tilde{\mathcal{F}}[r], s_r(G_n) \to q} v^*(G_n)$. Let

$$\lambda = \frac{\varepsilon/3}{M - b\big(\tilde{m}_r(q)\big) + \varepsilon/3}. \tag{3.9}$$

Let H denote the homothetic image of $D(\mathcal{F}_r)$ with center q and ratio λ. $D(\mathcal{F}_r)$ is a convex polyhedron, therefore in the topological sense, H is a neighborhood of q in the space $D(\mathcal{F}_r)$. As q is an accumulation point of $\tilde{S}_r$, and b is monotone and continuous, therefore there exists a graph $\tilde{G} \in \tilde{\mathcal{F}}[r]$ so that $s_r(\tilde{G}) \in H$, and

$$b\big(v^*(\tilde{G})\big) \ge b\big(\tilde{m}_r(q)\big) - \frac{2}{3}\varepsilon. \tag{3.10}$$

Denote the homothetic preimage of $s_r(\tilde{G})$ by q_0. Clearly, $s_r(\tilde{G}) = (1-\lambda) \cdot s_r(G) + \lambda \cdot q_0$.

Let $l = l_{s_r(\tilde{G})}$. Using Lemma 3.3, $u\big(l, s_r(\tilde{G})\big) = (1-\lambda) \cdot u\big(l, s_r(G)\big) + \lambda \cdot u(l, q_0)$, so

$$\begin{aligned} u\big(l, s_r(G)\big) &= \frac{1}{1+\lambda} u\big(l, s_r(\tilde{G})\big) - \frac{\lambda}{1+\lambda} u(l, q_0) \overset{(3.8)}{\ge} \frac{1}{1+\lambda} b\Big(m_r\big(s_r(\tilde{G})\big)\Big) - \frac{\lambda}{1+\lambda} M \\ &\overset{(3.5)}{\ge} \frac{1}{1+\lambda} b\big(v^*(\tilde{G})\big) - \frac{\lambda}{1+\lambda} M \overset{(3.10)}{\ge} \frac{1}{1+\lambda}\Big(b\big(\tilde{m}_r(q)\big) - \frac{2\varepsilon}{3}\Big) - \frac{\lambda}{1+\lambda} M \\ &= b\big(\tilde{m}_r(q)\big) - \frac{2\varepsilon}{3} - \frac{\lambda}{1+\lambda}\Big(M - b\big(\tilde{m}_r(q)\big) + \frac{2\varepsilon}{3}\Big) \overset{(3.9)}{=} b\big(\tilde{m}_r(q)\big) - \varepsilon. \end{aligned}$$

Let us use Lemma 3.8 with $X = S_r$ and $f(x) = b\big(m_r(x)\big) - \varepsilon$ and $f_x(y) = u(l_x, y)$. These satisfy the conditions of the lemma. Consider the sequences (x_i) and (λ_i) we get. Let g be a pair (X, g_X), where $P(X = x) = \lambda_i$ (and $X \notin (x_i)$ is impossible) and g_X is chosen with the

same distribution as with l_{x_i}. Let $\bar{l}\big[G,\omega,(X,g_X)\big] = l_X[G,\omega,g_X]$. This is a MRLA with radius r, satisfying that

$$\mathsf{E}_{\omega,g}\big(\bar{v}(G,\bar{l})\big) = \sum_i \lambda_i \mathsf{E}_{\omega,g_{x_i}}\big(\bar{v}(G,l_{x_i})\big) \overset{(3.7)}{=} \sum_i \lambda_i u\big(l_{x_i}, s_r(G)\big) \overset{(3.6)}{>} b\Big(m_r\big(s_r(G)\big)\Big) - \varepsilon$$

$$\overset{(3.5)}{\geq} b\big(v^*(G)\big) - \varepsilon. \qquad \square$$

Proof of Theorem 3.1. We get the statement from Theorem 3.2 with $b(x) = x - \varepsilon$. $\square$

3.3 The relations between preprocessing, mixing and randomizing

We show that preprocessing and mixing are equally strong tools. On one hand, Theorem 3.1 showed that MRLA and PMRLA are equally strong. On the other hand, it is easy to see the following theorem.

Lemma 3.9. *If a local choice problem is approximable by PMRLA, then this is approximable by PRLA, as well.*

Or in a more general form,

Lemma 3.10. *For each correct PMRLA l_1, there exists a correct PRLA l_2 using the same radius r and the same distribution of ω so that for each graph G,*

$$\mathsf{E}_{\omega}\Big(\bar{v}\big(G, l_2[G,\omega]\big)\Big) \geq \mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l_1[G,\omega,g]\big)\Big).$$

Proof. Roughly, we just change the global random seed g to the best seed depending on G. We show this idea in more detail.

If we use l_1 replacing g with a fixed value g_0, then we get a PRLA. Let us denote it by $l_1[g_0]$.

$$\mathsf{E}_{\omega}\Big(\bar{v}\big(G, l_1[g_0][G,\omega]\big)\Big) \tag{3.11}$$

is a function of g_0. For each graph G, let us choose such a $g_0 = g_0(G)$ for which (3.11) with $g_0 = g$ is at least as much as its expected value with respect to g. Let us define the PRLA $l_2 = l_1\big(g_0(G)\big)$. Then,

$$\mathsf{E}_{\omega,g}\Big(\bar{v}\big(G, l_1[G,\omega,g]\big)\Big) = \mathsf{E}_g\Big(\mathsf{E}_{\omega}\Big(\bar{v}\big(G, l_1[g][G,\omega]\big)\Big)\Big) \leq \mathsf{E}_{\omega}\Big(\bar{v}\Big(G, l_1\big[g_0(G)\big][G,\omega]\Big)\Big)$$

$$= \mathsf{E}_{\omega}\Big(\bar{v}\big(G, l_2[G,\omega]\big)\Big). \qquad \square$$

This means that preprocessing and mixing are interchangeable: either or both of them are equally strong. Now, we need to focus only on the strengths of mixing and randomizing. Obviously, every LA is a MLA and a RLA, furthermore every MLA and every RLA is a MRLA. There is no further relation between their strength: we show two problems, one is approximable by RLA, but not by MLA, and another one which is approximable by MLA but not by RLA.

An example for the former one follows from the results by Nguyen and Onak in [50]. They showed an approximating RLA for several problems such as the vertex cover. However, if

we have a transitive graph, say a cycle, then a MLA cannot distinguish between the vertices, therefore, for any fixed g, the algorithm either chooses all or none of the vertices. The latter solution is not a vertex cover, while former solution has relative size 1, which is far from optimal. Analogous argument works for matchings, as well.

The more surprising observation is that mixing can be useful when random labelling of all vertices does not solve the problem. We will use the following problem that we used in Chapter 2

Minimum Cut Problem. $\mathcal{F}$ is the family of all graphs that each vertex has at most 2 loops and at most d further edges going to other nodes. The number of loops are to express whether we consider the vertex a source (0), regular (1) or target (2). $C = \{0, 1, ...d, \text{“}T\text{”}\}$, where $c(x) = k$ expresses that x is on the source side with k neighbors in the target side, and $c(x) =$ “T” means that x is in the target side. $\delta = 1$, and $\mathcal{A}$ is defined as follows. For the root x of $\mathcal{F}_1$, if $c(x) \neq$ “T”, then x must have exactly $c(x)$ neighbors y (with multiplicity) with $c(y) =$ “T”. If x has no loop, the choice “T” is not allowed, and if x has two loops, the only allowed choice is “T”. $v($“T”$) = 0$ and $v(k) = -k$, expressing the negative of the size of the cut (which is to make it a maximization problem).

Lemma 3.11. *There is an approximate MLA for the Minimum Cut Problem.*

Proof. We use the local cut algorithm we presented in Theorem 2.2 for our special network with capacities 1 of all edges in both directions. □

For a graph G on $2k$ vertices, let $cut(G)$ denote the minimum number of edges between $X \subset V(G)$ and $V(G) - X$ with $|X| = k$. Let $C(d)$ denote the limsup of $cut(G)/\big|V(G)\big|$ on d-regular graphs G. Expander graphs show that $C(d)$ is positive.

Theorem 3.12. *For all RLA, there exists a graph G that for the choice c the RLA produces, $v^*(G) - \bar{v}(G, c)$ is arbitrary close to $C(d)$.*

Proof. Let r denote the radius the RLA uses. Consider a d-regular expander graph G on n vertices such that n is large enough and each subset of the vertices of size $\frac{n}{2} \pm o(n)$ cuts at least $C(d)n - o(n)$ edges. Let us start with no source and no target. Consider the expected propotion $\mathsf{E}_\omega\big(c^{-1}(\text{“}T\text{”})/n\big)$ of vertices chosen to the target side. Let us change some nodes one by one either to source or target nodes, if the ratio is more or less than $\frac{1}{2}$, respectively. Changing one node can change the choices only in the r-neighborhood of it, therefore, only in at most $(d+1)^r$ nodes. Furthermore, if we change all nodes, then the expected ratio $\mathsf{E}_\omega\big(c^{-1}(\text{“}T\text{”})/n\big)$ decreases to 0 or increases to 1, respectively. Therefore, we can stop the procedure at a point when this expected propotion is $\frac{1}{2} \pm o(1)$.

The choice at each node is independent from all but at most $(d+1)^{2r}$ nodes, therefore

$$Var\Big(\frac{c^{-1}(\text{“}T\text{”})}{n}\Big) = \frac{1}{n^2} \sum_{x,y \in V(G)} Cov\Big(c(x) = T, c(y) = T\Big) \leq \frac{1}{n^2} \sum_{x \in V(G)} (d+1)^{2r} = \frac{(d+1)^{2r}}{n}.$$

This implies that $c^{-1}(\text{“}T\text{”})/n = \frac{1}{2} + o(1)$ with high probability. Therefore, the expected size of the cut is $C(d)n - o(n)$.

Applying this observation to the Minimum Cut Problem as defined above, $v^*(G) = 0$, while $\bar{v}(G, c)$ can be arbitrary close to $-C(d)$. □

Now we have shown that neither of MLA and RLA is stronger than the other one. Finally, we show that MRLA is strictly stronger than the “union” of MLA and RLA. If we add up the problems of vertex cover and minimum cut, namely we want to construct both of them and the value of this is the sum of the two values, then this is approximable by neither MLA nor RLA, but this is approximable by MRLA.

Chapter 4

An undecidability result on limits of sparse graphs

Given a set $\mathcal{B}$ of finite rooted graphs and a radius r as an input, we prove that it is undecidable to determine whether there exists a sequence (G_i) of finite bounded degree graphs such that the rooted r-radius neighborhood of a random node of G_i is isomorphic to a rooted graph in $\mathcal{B}$ with probability tending to 1. Our proof implies a similar result for the case where the sequence (G_i) is replaced by a unimodular random rooted graph.

About dense graphs, Hatami and Norine [32] showed the undecidability of the problem of determining the validity of a linear weak inequality between the densities. Namely, let $t_H(G)$ denote the probability that $|V(H)|$ random nodes of G span H. Then the undecidable question is, given the sequence of graphs $H_1, H_2, ..., H_n$ and real coefficients $\lambda_1, \lambda_2, ..., \lambda_n$ as input, whether $\sum \lambda_i \cdot t_{H_i}(G) \geq 0$ holds for all graphs G. This result shows that the closure of the set of these distributions does not form a nice set, in some sense.

For sparse graphs, the closure of the set of distributions is convex, so contrary to the dense case, linear inequalities are sufficient to completely describe it. In this chapter, we will show the undecidability whether this closure contains a point at which all densities of some given neighborhoods are 0, or there is a positive lower bound of the total relative frequency of these neighborhoods in all graphs.

We mention that Bulitko investigated similar problems, namely, he showed the undecidability of the following. Given a rooted graph G as an input, does there exist a finite graph such that the 1-neighborhood of each of its nodes is isomorphic to G? [13] The main difference between the result of Bulitko and of ours is that Bulitko dealt with all neighborhoods, while we deal with almost all neighborhoods, in a particular sense.

We also mention that Bulitko showed that finding a finite or finding an infinite graph with the above property are not equivalent [12]. This implies the analogous nonequivalence for our case (Problem $\tilde{\boldsymbol{I}}$), as well.

4.1 Notation and the Aldous–Lyons problem

In this chapter, **graph** means **finite or infinite** graph with at least one vertex and with degrees bounded by a fix constant. For a graph G, a node o of G and an integer r, the ***r*-neighborhood** of o is the rooted subgraph $B_r(G, o)$ spanned by all nodes of G at distances at most r from o, with the root at o. The r-neighborhood distribution of a finite graph G is the probability distribution of the r-neighborhood of a uniform random node of G, up to isomorphism. Consider a sequence (G_n) of finite graphs. If for all r, the distribution of the

r-neighborhoods of G_n converges in n, then we say that (G_n) is **convergent**. In this case, there exists a random graph with the same distribution of the r-neighborhoods of the root as these limit distributions of (G_n). We call this random graph the **limit** of (G_n) and we call a random graph which is the limit of some sequence of finite graphs a **limit random graph**.

We do not have any nice description of the set of all limit random graphs. However, as Aldous and Lyons pointed out [1], there is a hope that this coincides with the set of all unimodular random graphs, which is something better understood. For example, Gábor Elek showed that each unimodular random graph can be represented by a topological graphing [22] which means a set of measure-preserving equivalence relations on a topological space with a probability measure. Whether these two sets really coincide is still an open question.

We show that describing either the set of all limit random graphs or of the unimodular random graphs is algorithmically undecidable in some sense. We also show that these sets are undecidable in the same way. Namely, we present a set of questions which are undecidable on either set, but for each question, the two answers are the same. We hope that our result can provide insight to the Aldous-Lyons problem.

4.2 Results

Problem *L* (limit)**.** *Given a set $\mathcal{B}$ of finite rooted graphs and a radius r as input, does there exist a limit random graph so that the r-neighborhood of the root belongs a.s. to $\mathcal{B}$?*
Problem *U* (unimodular)**.** *Given a set $\mathcal{B}$ of finite rooted graphs and a radius r as input, does there exist a unimodular random graph so that the r-neighborhood of the root belongs a.s. to $\mathcal{B}$?*
Problem *I* (infinite)**.** *Given a set $\mathcal{B}$ of finite rooted graphs and a radius r as input, does there exist a (possibly infinite) graph so that the r-neighborhood of each node belongs to $\mathcal{B}$?*

For an input $(\mathcal{B}, r)$ denote the three answers by $L(\mathcal{B}, r)$, $U(\mathcal{B}, r)$ and $I(\mathcal{B}, r)$, respectively. We know from [5] and [1] that $L(\mathcal{B}, r) \Rightarrow U(\mathcal{B}, r) \Rightarrow I(\mathcal{B}, r)$. The first implication is derived from a kind of unimodular property of the finite graphs, while the second one holds because if in a unimodular random graph u, the root a.s. has a local property, then for a randomly chosen graph from u, a.s. all its vertices has this property.

Theorem 4.1. *Problems L, U and I are undecidable. Moreover, there exists a subset of inputs such that, for each input $(\mathcal{B}, r)$, the three answers $L(\mathcal{B}, r)$, $U(\mathcal{B}, r)$ and $I(\mathcal{B}, r)$ are the same, and the three problems are undecidable on this set of inputs.*

Colored graph means a graph with directed and colored edges. All phenomena we defined above can also be defined for colored graphs using a fixed finite color set. (In this case, incidence is considered and distance in graph is measured by the undirected way.) First, we prove the same for the following two problems.
Problem $\tilde{L}$. *Given a set $\tilde{\mathcal{B}}$ of finite rooted colored graphs as input, does there exist a limit random colored graph so that the 2-neighborhood of the root belongs a.s. to $\tilde{\mathcal{B}}$?*
Problem $\tilde{U}$. *Given a set $\tilde{\mathcal{B}}$ of finite rooted colored graphs as input, does there exist a unimodular random colored graph so that the 2-neighborhood of the root belongs a.s. to $\tilde{\mathcal{B}}$?*
Problem $\tilde{I}$. *Given a set $\tilde{\mathcal{B}}$ of finite rooted colored graphs as input, does there exist a (possibly infinite) graph so that the 2-neighborhood of each node belongs to $\tilde{\mathcal{B}}$?*

For an input $\tilde{\mathcal{B}}$, denote the three answers by $L(\tilde{\mathcal{B}})$, $U(\tilde{\mathcal{B}})$ and $I(\tilde{\mathcal{B}})$, respectively. $L(\tilde{\mathcal{B}}) \Rightarrow U(\tilde{\mathcal{B}}) \Rightarrow I(\tilde{\mathcal{B}})$ holds by the same reason as in the uncolored case.

Proposition 4.2. *There exists a subset of inputs such that, for each input $\tilde{\mathcal{B}}$, the three answers $L(\tilde{\mathcal{B}})$, $U(\tilde{\mathcal{B}})$ and $I(\tilde{\mathcal{B}})$ are the same, and the three problems $\tilde{L}$, $\tilde{U}$ and $\tilde{I}$ are undecidable on this set of inputs.*

Proof of Proposition 4.2. We will show the undecidability of these problems even if $\tilde{\mathcal{B}}$ contains graphs only with degrees bounded by 4.

We prove this undecidability by reduction from Wang's Tiling Problem [55] which is the following. We get a finite set of equal-sized square tiles with a color on each edge. We want to place one of these on each square of a regular square grid so that abutting edges of adjacent tiles have the same color. The tiles cannot be rotated or reflected, but each tile can be used arbitrary many times, or in other words, we have infinitely many tiles of each kind in the set. The question is, given the set of tiles as input, whether there exists a tiling of the whole plane. Berger showed in 1966 that this question is algoritmically undecidable [6].

For each specific input of the tiling problem, we construct a set $\tilde{\mathcal{B}}$ such that the existence of a tiling of the plane is equivalent to all of $L(\tilde{\mathcal{B}})$, $U(\tilde{\mathcal{B}})$ and $I(\tilde{\mathcal{B}})$. This will imply the undecidabilities.

To each tiling of a part of the plane, we assign a graph, as follows. The nodes of this graph represent the different tiles. The color set of the edges is $\{\rightarrow, \uparrow\} \times$ (all colors in the tiles). If a tile Y is the right or up neighbor of X, and the color of the abutting edges is c then we put a directed edge from X to Y with color $(\rightarrow, c)$ or $(\uparrow, c)$, respectively.

Consider graphs corresponding to all 5×5 tilings. We define $\tilde{\mathcal{B}}$ as the set of 2-neighborhoods of the central nodes of all such graphs. Figure 1/b illustrates such a neighborhood, corresponding to the tiling in Figure 1/a.

Now we prove the equivalence of $L(\tilde{\mathcal{B}})$, $U(\tilde{\mathcal{B}})$, $I(\tilde{\mathcal{B}})$ and the existence of a tiling of the plane, by the following implications.

- $L(\tilde{\mathcal{B}}) \Rightarrow U(\tilde{\mathcal{B}}) \Rightarrow I(\tilde{\mathcal{B}})$, as we have already seen.
- If $I(\tilde{\mathcal{B}})$, then there exists a tiling of the plane.
  Proof. Consider such a graph. Each of its vertices can be represented by a tile, namely, the four colors of the tile are the four colors of the incident edges in the appropriate directions. Consider the lattice group, which is the group generated by two elements, called *up* and *right*, with the defining relation $up \cdot right = right \cdot up$. This group acts on the graph, because the structure of the neighborhoods guarantees that for each vertex v, $up\big(up^{-1}(v)\big) = right\big(right^{-1}(v)\big) = v$, and $up\big(right(v)\big) = right\big(up(v)\big)$. Let us take an arbitrary vertex v. Let us naturally identify the unit squares of the grid with the group elements. Then let the tile at each square s be the representing tile of $s(v)$. This is clearly a valid tiling.
- If there exists a tiling of the plane, then $L(\tilde{\mathcal{B}})$.
  Proof. Consider a tiling and let G_n be the graph corresponding to an $n \times n$ tiling. For each node in distance at least 2 from the border of the square, the 2-neighborhood of it is in $\tilde{\mathcal{B}}$. Thus the relative frequency of these nodes tends to 1, so the limit of the sequence (G_n) satisfies the requirements. □

Proof of Theorem 4.1. We construct a recursive function F that maps every finite set $\tilde{\mathcal{B}}$ of finite rooted colored graphs to a pair $(\mathcal{B}, r)$ so that $L(\mathcal{B}, r)$, $U(\mathcal{B}, r)$, $I(\mathcal{B}, r)$, $L(\tilde{\mathcal{B}})$, $U(\tilde{\mathcal{B}})$, $I(\tilde{\mathcal{B}})$ are all equivalent. This will prove the theorem because if there were an algorithm computing $L(\mathcal{B}, r)$, then we could compute $L(\tilde{\mathcal{B}})$, as well, by transforming $\tilde{\mathcal{B}}$ to $(\mathcal{B}, r)$ and then computing $L(\mathcal{B}, r)$.

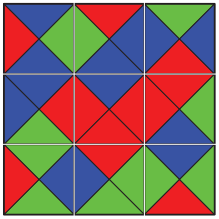


Figure 1/a

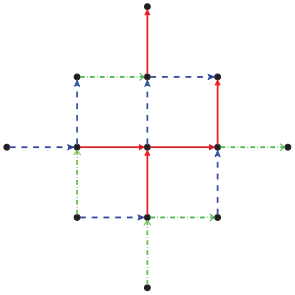


Figure 1/b

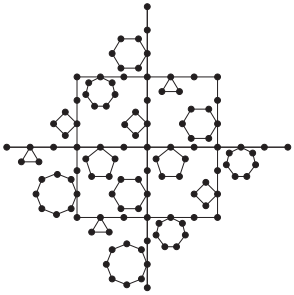


Figure 1/c

For the sake of simplicity, we assume that the colors in $\tilde{\mathcal{B}}$ are $\{1, 2, ..., k\}$. Take a 3-length path as a graph, let A, B, C and D be the four consecutive nodes of it. Take an l-length cycle, and identify one node of the cycle with B. If $l = 2c+1$ or $l = 2c+2$, then let us call this graph R_c and U_c, respectively. From a colored graph $\tilde{G}$, we construct a graph $G = f(\tilde{G})$ with degrees bounded by 4, like from Figure 1/b to Figure 1/c, as follows. Instead of each directed edge from X to Y with color $(\rightarrow, c)$ or $(\uparrow, c)$, we add a new copy of R_c or U_c, respectively, disjointly to the graph, and identify X with A and Y with D.

Let $\mathcal{G}$, $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{G}}_k$ denote the isomorphism classes of graphs, colored graphs and colored graphs with color set $\{1, 2, ..., k\}$, respectively. Let $\mathcal{G}_*$, $\tilde{\mathcal{G}}_*$ and $(\tilde{\mathcal{G}}_k)_*$ denote their rooted versions.

We will use a global retransformation $f^{-1} : \mathcal{G} \to \tilde{\mathcal{G}} \cup \{error\}$ and for each $k \in \mathbb{N}$, a local retransformation algorithm $f_k^{-1} : \mathcal{G}_* \to (\tilde{\mathcal{G}}_k)_* \cup \{error\}$, satisfying the following properties. (Note that f^{-1} denotes *not* the inverse of f, but something very similar.)

We define $f^{-1}(G)$ as follows. We call a node **central** if it has degree 4 and it has two neighbors with degree 4. These nodes will be the vertices of $f^{-1}(G)$. In G, let us cut these nodes into 4 nodes with degrees 1, keeping all edges. Then each component should be an R_i or U_i – otherwise we output an *error* – and we add the corresponding directed colored edge between the corresponding nodes of $f^{-1}(G)$. The colored graph we get is $f^{-1}(G)$.

We define $f_k^{-1}(G_*) = f_k^{-1}(G, o)$ as follows. We find the central node $\tilde{o}$ closest to o. This should be unique and in distance at most $k+2$ from o. $\tilde{o}$ will be the root of the colored graph. We take all nodes with degree 4 in the 6-neighborhood of $\tilde{o}$ as the vertices of the colored graph. We cut these nodes into 4 nodes with degrees 1, keeping all edges. We should get components with diameter at most $k+3$. We decode them into colored directed edges between the nodes. If everything was right then we output the 2-neighborhood of the root in this colored graph, otherwise we output an *error*.

1. $f^{-1}\big(f(\tilde{G})\big) = \tilde{G}$
2. f_k^{-1} depends only on the constant radius neighborhood of the root, namely,
$$f_k^{-1}(G_*) = f_k^{-1}\big(B_{2k+8}(G_*)\big)$$
3. If $f_k^{-1}(G_*) \in (\tilde{\mathcal{G}}_k)_*$, then each vertex of $f_k^{-1}(G_*)$ is in distance at most 2 from the root.
4. A graph is globally retransformable if and only if it is locally retransformable at each point. Namely,
$$\big(\forall o \in V(G) : f_k^{-1}(G, o) \in (\tilde{\mathcal{G}}_k)_*\big) \Leftrightarrow \big(f^{-1}(G) \in \tilde{\mathcal{G}}_k\big).$$

5. There exists a constant C_k that if $\tilde{G} = f^{-1}(G) \neq error$, then there exists a mapping $m : V(G) \to V(\tilde{G})$ satisfying the followings.

$$\forall v \in V(G) : \ B_2\big(m(v)\big) \cong f_k^{-1}(G, v)$$

$$\forall \tilde{v} \in V(\tilde{G}) : \ 1 \leq \big|m^{-1}(\tilde{v})\big| \leq C_k$$

Properties 1, 2, 3 and 4 are obvious consequences of the retransformating methods. At property 5, $m(v)$ can be constructed as follows. We take the central node v_0 closest to v. In the definition of f^{-1}, the new nodes are constructed from the central nodes of the original graph. Let $m(v)$ be the node of $\tilde{G}$ corresponding to v_0. Then $B_2\big(m(v)\big) \cong f_k^{-1}(G, v)$, by the definition of f_k^{-1}. About $m^{-1}(\tilde{v})$, it contains the central node of G corresponding to $\tilde{v}$, and this node must be in distance at most $k + 2$ from each node in $m^{-1}(\tilde{v})$, according to the construction of m. That is why, $\big|m^{-1}(\tilde{v})\big| \leq 1 + 4 + 4^2 + ... + 4^{k+2} = C_k$.

Let $\mathcal{B} = F(\tilde{\mathcal{B}})$ be the set of rooted graphs G_* satisfying that its degrees are at most 4, and its nodes are in distance at most $r = 2k + 8$ from the root, and $f_k^{-1}(G_*) \in \tilde{\mathcal{B}}$. Clearly, F is a recursive function.

If $L(\tilde{\mathcal{B}})$, then there exists a finite colored graph sequence $(\tilde{G}_n)$ with the relative frequency of the nodes with 2-neighborhoods from $\tilde{\mathcal{B}}$ tending to 1. Then using Property 5, we get that the relative frequency of r-neighborhoods of $f(\tilde{G}_n)$ which are from $\mathcal{B}$ tends to 1. This implies $L(\mathcal{B}, r)$.

If $I(\mathcal{B}, r)$, then take a graph G so that $\forall v \in V(G) : B_r(v) \in \mathcal{B}$. By Properties 1 and 4, $f^{-1}(G) \neq error$, and $\forall v \in V(G) : B_2\big(f_r^{-1}(G, v)\big) \in \tilde{\mathcal{B}}$. Thus, $f^{-1}(G)$ shows that $I(\tilde{\mathcal{B}})$.

To sum up, $L(\mathcal{B}, r) \Rightarrow U(\mathcal{B}, r) \Rightarrow I(\mathcal{B}, r) \Rightarrow I(\tilde{\mathcal{B}}) \Leftrightarrow U(\tilde{\mathcal{B}}) \Leftrightarrow L(\tilde{\mathcal{B}}) \Rightarrow L(\mathcal{B}, r)$, which proves the equivalences. □

Chapter 5

On the independence ratio of regular graphs with large girth

An *independent set* is a set of vertices in a graph, no two of which are adjacent. The *independence ratio* of a graph is the size of its largest independent set divided by the total number of vertices. Let $d \geq 3$ be an integer and suppose that G is a d-regular finite graph with sufficiently large girth, that is, G does not contain cycles shorter than a sufficiently large given length. In other words, G locally looks like a d-regular tree. What can we say about the independence ratio of G? In a regular (infinite) tree "every other vertex" can be chosen, so one is tempted to say that the independence ratio should tend to $1/2$ when the girth goes to infinity. This is not the case, however. Bollobás showed that the independence ratio of random d-regular graphs is separated from $1/2$ for any given d [7]. In [48] McKay sharpened the work of Bollobás and obtained somewhat better estimates. For instance, he proved that the independence ratio of a random 3-regular graph is at most 0.45537 (with high probability).

Actually, a random regular graph does not have large girth, but it can be easily altered into a regular graph with large girth and with essentially the same independence ratio. Therefore any lower bound for the independence ratio of regular graphs with large girth applies to random regular graphs as well. We now briefly survey what is known about the independence ratio of regular graphs with large girth. According to a result of Shearer [51] for any triangle-free graph with average degree d the independence ratio is at least

$$\frac{d \log d - d + 1}{(d-1)^2}.$$

Even though this bound is true in a very general setting, asymptotically speaking (as $d \to \infty$) it is as good as any known lower bound. Shearer himself found an improvement for (regular) large girth graphs [52]. Lauer and Wolmard further improved that bound for regular, large girth graphs with degree $d \geq 7$ [41]. However, asymptotically, all of these bounds are the same: $(\log d)/d$. Note that the previously mentioned upper bounds of Bollobás and McKay are both asymptotically equal to $2(\log d)/d$. Finally, we mention results where computer assistance was needed to find/verify the obtained bounds. In his thesis [34] Hoppen presents an approach that outdoes the above-mentioned bounds when the degree is small ($d \leq 10$). For $d = 3$ Kardoš, Král and Volec improved Hoppen's method and obtained the bound 0.4352 [39]. For $d = 3$, the highest possible ratio achievable by a local algorithm is certainly smaller than McKay's upper bound 0.45537 on the independence ratio of random 3-regular graphs.

The main result of this paper is the next theorem.

Theorem 5.1. *Every 3-regular graph with sufficiently large girth has independence ratio at least* 0.4361.

Next we briefly outline the proof of Theorem 5.1. Suppose that there are real numbers assigned to each vertex of G. We always get an independent set by choosing those vertices having larger number than each of their neighbors. If we assign these numbers to the vertices in some random manner, then we get a random independent set. If the expected size of this random independent set can be computed, then it gives a lower bound on the independence ratio. In many cases, the probability that a given vertex is chosen is the same for all vertices, whence this probability itself is a lower bound.

The idea is to consider a random assignment that is *almost* an eigenvector (with high probability) with some negative eigenvalue λ. Then we expect many of the vertices with positive numbers to be larger than their neighbors. The spectrum of the d-regular (infinite) tree is $[-2\sqrt{d-1}, 2\sqrt{d-1}]$, so it is reasonable to expect that we can find such a random assignment for $\lambda = -2\sqrt{d-1}$. As we will see, the approach described above can indeed be carried out, and it produces a lower bound 0.4298 in the $d = 3$ case. This is surprisingly good given that it comes from a fairly simple random procedure and that we do not need computer to determine the bound. As far as the authors know, the best bound obtained without the use of computers is 0.4139 and is due to Shearer. See [41, Table 1].

Moreover, this construction provides two disjoint indepentent set of this size, which is a bipartite graph spanned by 0.85965 ratio of the vertices. Before this result, the best known bound was less than 0.818 [33]. Moreover, this was not a mathematically proved ratio, but it was the result of a statistical calculation by a computer program on large random cubic graphs.

Using the same random assignment but a more sophisticated way to choose the vertices for our independent set provides a better bound. We fix some threshold $\tau \in \mathbb{R}$, and, like in a percolation, we only keep those vertices that are above this threshold. We choose τ in such a way that the components of the remaining vertices are small with high probability. We omit the large components and we choose an independent set from each of the small components. (Note that the small components are all trees provided that the girth of the original graph is large enough. Since trees are bipartite, they have an independent set containing at least half of the vertices.) With the right choice of τ this method yields a better bound than the original approach. However, it seems very hard to actually compute this bound. We simulated the above random procedure by computer and it suggested that the probability for any given vertex to be in the independent set is above 0.438 in the 3-regular case. The best bound that we can prove rigorously is 0.4361.

Random processes on the regular tree

In fact, instead of working on finite graphs with large girth, it will be more convenient for us to consider the regular (infinite) tree and look for independent sets on this tree that are i.i.d. factors.

Let T_d denote the d-regular tree for some positive integer $d \geq 3$, $V(T_d)$ is the vertex set, and $\mathrm{Aut}(T_d)$ is the group of graph automorphisms of T_d. Suppose that we have independent standard normal random variables Z_v assigned to each vertex $v \in V(T_d)$. We call an instance of an assignment a configuration. A *factor of i.i.d. independent set* is a random independent set that is obtained as a measurable function of the configuration and that commutes with the natural action of $\mathrm{Aut}(G)$. By a *factor of i.i.d. process* we mean random variables X_v, $v \in V(T_d)$ that are all obtained as measurable functions of the random variables Z_v and that are $\mathrm{Aut}(T_d)$-invariant. Actually, in this paper we will only consider *linear factor of i.i.d. processes* defined

as follows.

Definition 5.2. We say that a process X_v, $v \in V(T_d)$ is a *linear factor* of the i.i.d. process Z_v if there exist real numbers $\alpha_0, \alpha_1, \ldots$ such that

$$X_v = \sum_{u \in V(T_d)} \alpha_{d(v,u)} Z_u = \sum_{k=0}^{\infty} \sum_{u : d(v,u)=k} \alpha_k Z_u, \tag{5.1}$$

where $d(v,u)$ denotes the distance between the vertices v and u in T_d. Note that the infinite sum in (5.1) converges almost surely if and only if $\alpha_0^2 + \sum_{k=1}^{\infty} d(d-1)^{k-1}\alpha_k^2 < \infty$.

These linear factors are clearly $\mathrm{Aut}(T_d)$-invariant. Furthermore, the random variable X_v defined in (5.1) is always a centered Gaussian.

Definition 5.3. We call a collection of random variables X_v, $v \in V(T_d)$ a *Gaussian process on* T_d if they are jointly Gaussian and each X_v is centered. (Random variables are said to be jointly Gaussian if any finite linear combination of them is Gaussian.)

Furthermore, we say that a Gaussian process X_v is invariant if it is $\mathrm{Aut}(T_d)$-invariant, that is, for arbitrary graph automorphism $\Phi : V(T_d) \to V(T_d)$ of T_d the joint distribution of the Gaussian process $X_{\Phi(v)}$ is the same as that of the original process.

The following invariant processes will be of special interest for us.

Theorem 5.4. *For any real number λ with $|\lambda| \le d$, there exists a non-trivial invariant Gaussian process X_v on T_d that satisfies the eigenvector equation with eigenvalue λ, i.e., (with probability 1) for every vertex v, it holds that*

$$\sum_{u \in N(v)} X_u = \lambda X_v,$$

where $N(u)$ denotes the set of neighbors of v.

The joint distribution of such a process is unique under the additional condition that the variance of X_v is 1. We will refer to this (essentially unique) process as the Gaussian wave function *with eigenvalue λ.*

These Gaussian wave functions can be approximated by linear factor of i.i.d. processes provided that $|\lambda| \le 2\sqrt{d-1}$.

Theorem 5.5. *For any real number λ with $|\lambda| \le 2\sqrt{d-1}$, there exist linear factor of i.i.d. processes that converge in distribution to the Gaussian wave function corresponding to λ.*

The rest of the paper is organized as follows: in Section 5.1 we prove Theorem 5.4 and derive some useful properties of Gaussian wave functions, in Section 5.2 we give a proof for Theorem 5.5, and in Section 5.3 we show how one can use these random processes to find large independent sets.

5.1 Invariant Gaussian wave functions

We call the random variables X_v, $v \in V(T_d)$ a Gaussian process if they are jointly Gaussian and each X_v is centered (see Definition 5.3). The joint distribution is completely determined

by the covariances $\mathrm{cov}(X_u, X_v)$, $u, v \in V(T_d)$. A Gaussian process with prescribed covariances exists if and only if the corresponding infinite "covariance matrix" is positive semidefinite.

From this point on, all the Gaussian processes considered will be $\mathrm{Aut}(T_d)$-invariant. For an invariant Gaussian process X_v the covariance $\mathrm{cov}(X_u, X_v)$ clearly depends only on the distance $d(u, v)$ of u and v. (The distance between the vertices u, v is the length of the shortest path connecting u and v in T_d.) Let us denote the covariance corresponding to distance d by σ_d. So an invariant Gaussian process is determined (in distribution) by the sequence $\sigma_0, \sigma_1, \dots$ of covariances.

Theorem 5.4 claims that for any $|\lambda| \le d$ there exists an invariant Gaussian process that satisfies the eigenvector equation $\sum_{u \in N(v)} X_u = \lambda X_v$ for each vertex v. What would the covariance sequence of such a *Gaussian wave function* be? Let $v_1, \dots, v_d$ denote the neighbors of an arbitrary vertex v_0. Then

$$0 = \mathrm{cov}\,(X_{v_0}, 0) = \mathrm{cov}\,(X_{v_0}, X_{v_1} + \cdots + X_{v_d} - \lambda X_{v_0}) = d\sigma_1 - \lambda\sigma_0.$$

Also, if u is at distance k from v_0, then it has distance $k-1$ from one of the neighbors $v_1, \dots, v_d$, and has distance $k+1$ from the remaining $d-1$ neighbors of v_0. Therefore

$$0 = \mathrm{cov}\,(X_{v_0}, 0) = \mathrm{cov}\,(X_u, X_{v_1} + \cdots + X_{v_d} - \lambda X_{v_0}) = (d-1)\sigma_{k+1} + \sigma_{k-1} - \lambda\sigma_k.$$

After multiplying our process with a constant we may assume that the variance of X_v is 1, that is, $\sigma_0 = 1$. So the covariances satisfy the following linear recurrence relation:

$$\sigma_0 = 1;\ d\sigma_1 - \lambda\sigma_0 = 0;\ (d-1)\sigma_{k+1} - \lambda\sigma_k + \sigma_{k-1} = 0,\ k \ge 1. \tag{5.2}$$

There is a unique sequence σ_k satisfying the above recurrence. Therefore to prove the existence of the Gaussian wave function we only need to check that the corresponding infinite matrix is positive semidefinite. This does not seem to be a straightforward task, though, so we take another approach instead, where we recursively consruct the Gaussian wave function. (This approach will also yield some interesting and useful properties of these Gaussian wave functions, see Remark 5.7 and 5.8.)

Remark 5.6. The case $|\lambda| \le 2\sqrt{d-1}$ also follows from the results presented in the next section, where we construct factor of i.i.d. processes, the covariance matrices of which converge to the "covariance matrix" of the (supposed) Gaussian wave function. As the limit of positive semidefinite matrices, this "covariance matrix" is positive semidefinite, too, and thus the Gaussian wave function indeed exists.

Proof of Theorem 5.4. Let σ_k be the solution of the recurrence relation (5.2), in particular,

$$\sigma_0 = 1;\ \sigma_1 = \frac{\lambda}{d};\ \sigma_2 = \frac{\lambda^2 - d}{d(d-1)}.$$

We need to find a Gaussian process X_v, $v \in V(T_d)$ such that

$$\mathrm{cov}(X_u, X_v) = \sigma_{d(u,v)} \tag{5.3}$$

holds for all $u, v \in V(T_d)$.

We will define the random variables X_v recursively on larger and larger connected subgraphs of T_d. Suppose that the random variables X_v are already defined for $v \in H$ such that (5.3) is satisfied for any $u, v \in H$, where H is a (finite) set of vertices for which the induced subgraph $T_d[H]$ is connected. Let v_0 be a leaf (i.e., a vertex with degree 1) in $T_d[H]$, v_d denotes the

unique neighbor of v_0 in $T_d[H]$, and $v_1, \ldots, v_{d-1}$ are the remaining neighbors in T_d. We now define the random variables $X_{v_1}, \ldots, X_{v_{d-1}}$. Let $(Y_1, \ldots, Y_{d-1})$ be a multivariate Gaussian that is independent from X_v, $v \in H$ and that has a prescribed covariance matrix that we will specify later. Set

$$X_{v_i} \stackrel{\text{def}}{=} \frac{\lambda}{d-1} X_{v_0} - \frac{1}{d-1} X_{v_d} + Y_i \ , \ i = 1, \ldots, d-1.$$

For $1 \le i \le d-1$ we have

$$\operatorname{cov}(X_{v_i}, X_{v_0}) = \frac{\lambda}{d-1} - \frac{1}{d-1}\sigma_1 = \sigma_1,$$

and if $u \in H \setminus \{v_0\}$ is at distance $k \ge 1$ from x_0, then

$$\operatorname{cov}(X_{v_i}, X_u) = \frac{\lambda}{d-1}\sigma_k - \frac{1}{d-1}\sigma_{k-1} = \sigma_{k+1}.$$

We also need that

$$\operatorname{var}(X_{v_i}) = \sigma_0 \ \text{ and } \ \operatorname{cov}\left(X_{v_i}, X_{v_j}\right) = \sigma_2, \text{ whenever } 1 \le i, j \le d-1, \ i \ne j. \tag{5.4}$$

Since

$$\operatorname{var}(X_{v_i}) = \left(\frac{\lambda}{d-1}\right)^2 + \left(\frac{1}{d-1}\right)^2 - \frac{2\lambda}{(d-1)^2}\sigma_1 + \operatorname{var}(Y_i) \text{ and}$$
$$\operatorname{cov}\left(X_{v_i}, X_{v_j}\right) = \left(\frac{\lambda}{d-1}\right)^2 + \left(\frac{1}{d-1}\right)^2 - \frac{2\lambda}{(d-1)^2}\sigma_1 + \operatorname{cov}(Y_i, Y_j),$$

we can set $\operatorname{var}(Y_i)$ and $\operatorname{cov}(Y_i, Y_j)$ such that (5.4) is satisfied, namely let

$$\operatorname{var}(Y_i) = a \stackrel{\text{def}}{=} \frac{(d-2)(d^2-\lambda^2)}{d(d-1)^2} \text{ and } \operatorname{cov}(Y_i, Y_j) = b \stackrel{\text{def}}{=} \frac{-(d^2-\lambda^2)}{d(d-1)^2}.$$

We still have to show that there exist Gaussians $Y_1, \ldots, Y_{d-1}$ with the above covariances. The corresponding $(d-1) \times (d-1)$ covariance matrix would have a's in the main diagonal and b's everywhere else. The eigenvalues of this matrix are $a + (d-2)b$ and $a - b$ (with $a - b$ having multiplicity $d-2$). Therefore the matrix is positive semidefinite if $a \ge b$ and $a \ge -(d-2)b$. It is easy to check that these inequalities hold when $|\lambda| \le d$. (In fact, $a = -(d-2)b$, so the covariance matrix is singular, which means that there is some linear dependence between $Y_1, \ldots, Y_{d-1}$. Actually, this linear dependence is that $Y_1 + \cdots + Y_{d-1} = 0$, and that is why the eigenvector equation $X_{v_1} + \cdots + X_{v_d} = \lambda X_{v_0}$ holds.)

So the random variables X_v are now defined on the larger set $H' = H \cup \{v_1, \ldots, v_{d-1}\}$ such that (5.3) is satisfied for any $u, v \in H'$. Since

$$\begin{pmatrix} 1 & \sigma_1 \\ \sigma_1 & 1 \end{pmatrix}$$

is positive semidefinite for $|\lambda| \le d$, we can start with a set H containing two adjacent vertices, and then in each step we can add to H the remaining $d-1$ vertices of a leaf. □

Remark 5.7. There is an important consequence of the proof above, which we will make use of, when we will need to compute the probability of certain configurations for a particular Gaussian wave function in Section 5.3. Let u and v be adjacent vertices in T_d. They cut T_d (and thus the Gaussian wave function on it) into two parts. Our proof yields that the two parts of the process are independent under the condition $X_u = x_u; X_v = x_v$ for any real numbers x_u, x_v.

Remark 5.8. If $d = 3$ and $\lambda = -2\sqrt{2}$, then we have $Y_2 = -Y_1$ in the above proof with $\mathrm{var}(Y_1) = a = 1/12$. So we can express X_{v_1} and X_{v_2} with the standard Gaussian $Z = 2\sqrt{3}Y_2$ as follows:

$$X_{v_1} = -\sqrt{2}X_{v_0} - \frac{1}{2}X_{v_3} - \frac{1}{2\sqrt{3}}Z \text{ and}$$
$$X_{v_2} = -\sqrt{2}X_{v_0} - \frac{1}{2}X_{v_3} + \frac{1}{2\sqrt{3}}Z.$$

Note that Z is independent from the random variables $X_v, v \in H$, in particular, it is independent from X_{v_0}, X_{v_3}.

5.1.1 Correlated percolations corresponding to Gaussian wave functions

Let X_v, $v \in V(T_d)$ be some fixed invariant process on T_d. For any $\tau \in \mathbb{R}$ we define

$$H_\tau \stackrel{\text{def}}{=} \{v \in V(T_d) \ : \ X_v \geq \tau\},$$

that is, we throw away the vertices below some threshold τ. One very natural question about this random set H_τ is whether its components are finite almost surely or not. Clearly, there exists a *critical threshold* $\tau_c \in [-\infty, \infty]$ such that for $\tau > \tau_c$ the component of any given vertex is finite almost surely, while if $\tau < \tau_c$, then any given vertex has infinite component with some positive probability.

First we explain why it would be extremely useful for us to determine this critical threshold (or bound it from above). Let τ be above the critical threshold τ_c and let I_τ be the "largest" independent set contained by H_τ. More precisely, we choose the largest independent set in each of the (finite) components of H_τ and consider their union. If the largest independent set is not unique, then we choose one in some invariant way. This way we get an invariant independent set I_d. Since $H_{\tau'} \supseteq H_\tau$ if $\tau' < \tau$, for smaller τ we will get larger independent sets (i.e., the probability that a given vertex is in the independent set is larger). Therefore, in this construction, we want to pick τ close to the critical threshold.

The next lemma provides a sufficient condition for the components to be finite in the case when our process X_v is a Gaussian wave function. Let us fix path in T_d containing $m + 2$ vertices for some positive integer m and consider the event that the random variable assigned to each of the $m+2$ vertices is at least τ. The sufficient condition is roughly the following: for any $x, y \geq \tau$, the probability of the event described above is less than $1/(d-1)^m$ under the condition that the value assigned to the first and second vertex is x and y, respectively. In fact, the only thing that we will use about Gaussian wave functions is their property pointed out in Remark 5.7.

Lemma 5.9. *Let $X_v, v \in T_d$ be a Gaussian wave function on T_d and let $v_{-1}, v_0, v_1, \ldots, v_m$ be any fixed path in T_d containing $m + 2$ vertices for some positive integer m. Suppose that there exists a real number $c < 1/(d-1)^m$ such that*

$$P\left(X_{v_i} \geq \tau \ , \ 1 \leq i \leq m | X_{v_{-1}} = x_{v_{-1}}; X_{v_0} = x_{v_0}\right) < c$$

holds for any real numbers $x_{v_{-1}}, x_{v_0} \geq \tau$. Then each component of

$$H_\tau = \{v \in V(T_d) \ : \ X_v \geq \tau\} \subset V(T_d)$$

is finite almost surely.

Proof. Let u be an arbitrary vertex and let us consider the component of u in H_τ. The assumptions of the lemma clearly imply that the expected number of vertices in the component at distance $2sm+1$ from u is at most $d(d-1)^{sm}c^s$, which is exponentially small in s. The Markov inequality yields that the probability that the component has at least one vertex at distance $2sm+1$ is exponentially small, too. It follows that each component must be finite with probability 1. □

Using this criterion we can give an upper bound for the critical threshold in the case $d=3$, $\lambda=-2\sqrt{2}$.

Theorem 5.10. *Let $d=3$ and let X_v, $v\in T_3$ be the Gaussian wave function with eigenvalue $\lambda=-2\sqrt{2}$. If $\tau\geq -0.086$, then each component of H_τ is finite almost surely.*

Proof. We will use Lemma 5.9 with $m=2$. Let $\tau=-0.086$ and let us fix a path containing four vertices of T_3. We will denote the random variables assigned to the vertices X,Y,U,V. Let x,y be arbitrary real numbers not less than τ. From now on, every event and probability will be meant under the condition $X=x;Y=y$. According to Remark 5.8 there exist independent standard normal random variables Z_1,Z_2 such that

$$U=-\sqrt{2}y-\frac{1}{2}x-\frac{1}{2\sqrt{3}}Z_1;$$
$$V=-\sqrt{2}U-\frac{1}{2}y-\frac{1}{2\sqrt{3}}Z_2=\frac{3}{2}y+\frac{1}{\sqrt{2}}x+\frac{1}{\sqrt{6}}Z_1-\frac{1}{2\sqrt{3}}Z_2.$$

Our goal is to prove that the probability of $U\geq\tau;V\geq\tau$ is less than $1/4$ for any fixed $x,y\geq\tau$. If we decrease y by some positive Δ and increase x by $2\sqrt{2}\Delta$ at the same time, then U does not change, while V gets larger, whence the probability in question grows. Thus setting y equal to τ and changing x accordingly always yield a larger probability. So from now on we will assume that $y=\tau$. Then

$$U\geq\tau\Leftrightarrow Z_1\leq-\sqrt{3}x-2\sqrt{6}\tau-2\sqrt{3}\tau;$$
$$V\geq\tau\Leftrightarrow -Z_1+\frac{1}{\sqrt{2}}Z_2\leq\sqrt{3}x+\frac{\sqrt{3}}{\sqrt{2}}\tau.$$

We notice that the sum of the right hand sides is independent from x:

$$a\stackrel{\text{def}}{=}-\tau\left(2\sqrt{6}+2\sqrt{3}-\frac{\sqrt{3}}{\sqrt{2}}\right).$$

Therefore we have to maximize the following probability in d_1:

$$P\left(Z_1\leq d_1;Z_2/q\leq Z_1+a-d_1\right),\text{ where }q=\sqrt{2}.$$

This can be expressed as a two-dimensional integral:

$$f(d_1)\stackrel{\text{def}}{=}\int_{-\infty}^{d_1}\int_{-\infty}^{q(z_1+a-d_1)}\frac{1}{2\pi}\exp\left(-\frac{z_1^2+z_2^2}{2}\right)\mathrm{d}z_2\,\mathrm{d}z_1. \tag{5.5}$$

To find the maximum of the function $f(d_1)$, we take its derivative, which can be expressed using the cumulative distribution function Φ of the standard normal distribution:

$$f'(d_1)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-d_1^2}{2}\right)\Phi(qa)-\frac{\sqrt{1+q^2}}{\sqrt{2\pi}q}\exp\left(\frac{-d_2^2}{2}\right)\Phi\left(\sqrt{1+q^2}a-d_2/q\right),$$
$$\text{where }d_2=\frac{q}{\sqrt{1+q^2}}(a-d_1).$$

The derivative has a unique root, belonging to the maximum of f. Solving $f'(d_1) = 0$ numerically ($d_1 \approx 0.555487$), then computing the integral (5.5) (≈ 0.249958) shows that $\max f < 1/4$ as claimed. □

5.2 Approximation with factor of i.i.d. processes

Our goal in this section is to prove Theorem 5.5: there exist linear factor of i.i.d. processes approximating (in distribution) the Gaussian wave function with eigenvalue λ provided that $|\lambda| \le 2\sqrt{d-1}$. This will follow easily from the next lemma.

Lemma 5.11. *Let $|\lambda| \le 2\sqrt{d-1}$ be fixed. For a sequence of real numbers $\alpha_0, \alpha_1, \ldots$, we define the sequence $\delta_0, \delta_1, \ldots$ as*

$$\delta_0 \stackrel{def}{=} d\alpha_1 - \lambda\alpha_0;\ \delta_k \stackrel{def}{=} (d-1)\alpha_{k+1} - \lambda\alpha_k + \alpha_{k-1},\ k \ge 1. \tag{5.6}$$

Then for any $\varepsilon > 0$ there exists a sequence α_k such that

$$\alpha_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\alpha_k^2 = 1 \text{ and } \delta_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\delta_k^2 < \varepsilon.$$

We can clearly assume that only finitely many α_k are nonzero.

Remark 5.12. We can think of such sequences α_k as invariant approximate eigenvectors on T_d. Let us fix a root of T_d and write α_k on vertices at distance k from the root. Then the vector $f \in \ell_2(V(T_d))$ obtained is spherically symmetric around the root (i.e., f is invariant under automorphisms fixing the root). Furthermore, $\|f\|^2 = \alpha_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\alpha_k^2$.

As for the sequence δ_k, it corresponds to the vector $A_{T_d}f - \lambda f \in \ell_2(V(T_d))$, where A_{T_d} denotes the adjacency operator of T_d. Therefore $\|A_{T_d}f - \lambda f\|^2 = \delta_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\delta_k^2$. So the real content of the above lemma is that for any $\varepsilon > 0$ there exists a spherically symmetric vector $f \in \ell_2(V(T_d))$ such that $\|f\| = 1$ and $\|A_{T_d}f - \lambda f\| < \varepsilon$.

In the best scenario $\delta_k = 0$ would hold for each k, that is, α_k would satisfy the following linear recurrence:

$$r\alpha_1 - \lambda\alpha_0 = 0; (d-1)\alpha_{k+1} - \lambda\alpha_k + \alpha_{k-1} = 0,\ k \ge 1. \tag{5.7}$$

However, for a non-trivial solution α_k of the above recurrence we always have $\alpha_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\alpha_k^2 = \infty$. This follows from the fact that the point spectrum of A_{T_d} is empty.

First we show how Theorem 5.5 follows from the above lemma.

Proof of Theorem 5.5. Suppose that we have independent standard normal random variables Z_v. Let $\varepsilon > 0$ and let α_k as in Lemma 5.11. Let X_v be the linear factor of Z_v with coefficients α_k as in (5.1). Then

$$\mathrm{var}(X_v) = \alpha_0^2 + \sum_{k\ge 1} d(d-1)^{k-1}\alpha_k^2 = 1.$$

Let v_0 be an arbitrary vertex with neighbors $v_1, \ldots, v_d$. It is easy to see that

$$X_{v_1} + \ldots + X_{v_d} - \lambda X_{v_0} = (d\alpha_1 - \lambda\alpha_0)Z_{v_0} + \sum_{k=1}^{\infty} \sum_{u:d(v_0,u)=k} \left((d-1)\alpha_{k+1} - \lambda\alpha_k + \alpha_{k-1}\right) Z_u.$$

So $X_{v_1}+\ldots+X_{v_d}-\lambda X_{v_0}$ is also a linear factor with coefficients δ_k as defined in (5.6). Therefore the variance of $X_{v_1}+\ldots+X_{v_d}-\lambda X_{v_0}$ is $\delta_0^2+\sum_{k\geq 1} d(d-1)^{k-1}\delta_k^2<\varepsilon$.

What can we say about the covariance sequence σ_k of the Gaussian process X_v? We have $\sigma_0=1$ and

$$|d\sigma_1-\lambda\sigma_0|\,,|(d-1)\sigma_{k+1}-\lambda\sigma_k+\sigma_{k-1}|\leq\sqrt{\operatorname{var}(X_u)\operatorname{var}\left(X_{v_1}+\cdots+X_{v_d}-\lambda X_{v_0}\right)}<\sqrt{\varepsilon}.$$

In other words, the equations in (5.2) hold with some small error $\sqrt{\varepsilon}$. If K is a positive integer and $\delta>0$ is a real number, then for sufficiently small ε we can conclude that for $k\leq K$ the covariance σ_k is closer than δ to the actual solution of (5.2). It follows that if ε tends to 0, then the covariance sequence of X_v pointwise converges to the unique solution of (5.2). It follows that X_v converges to the Gaussian wave function in distribution as $\varepsilon\to 0$. □

Proof of Lemma 5.11. It is enough to prove the statement for $|\lambda|<2\sqrt{d-1}$, the case $\lambda=\pm 2\sqrt{d-1}$ then clearly follows. Excluding $\pm 2\sqrt{d-1}$ will spare us some technical difficulties.

Let β_k be a solution of the following recurrence

$$r\beta_1-\lambda\sqrt{d-1}\beta_0=0;\beta_{k+1}-\frac{\lambda}{\sqrt{d-1}}\beta_k+\beta_{k-1}=0,\ k\geq 1. \tag{5.8}$$

(This is the recurrence that we would get from (5.7) had we made the substitution $\beta_k=(d-1)^{k/2}\alpha_k$.) Since $|\lambda|<2\sqrt{d-1}$, the quadratic equation $x^2-\frac{\lambda}{\sqrt{d-1}}x+1=0$ has two complex roots, both of norm 1, which implies that (5.8) has bounded solutions. Set

$$\alpha_k\stackrel{\text{def}}{=}\varrho^k(d-1)^{-k/2}\beta_k \tag{5.9}$$

for some positive real number $1/2\leq\varrho<1$. Since β_k is bounded, $\alpha_0^2+\sum_{k\geq 1}d(d-1)^{k-1}\alpha_k^2$ is finite for any $\varrho<1$. As $\varrho\to 1-$, however, $\alpha_0^2+\sum_{k\geq 1}d(d-1)^{k-1}\alpha_k^2$ tends to infinity. Furthermore,

$$\delta_k=(d-1)\alpha_{k+1}-\lambda\alpha_k+\alpha_{k-1}=(d-1)^{-(k-1)/2}\varrho^k\left(\varrho\beta_{k+1}-\frac{\lambda}{\sqrt{d-1}}\beta_k+\varrho^{-1}\beta_{k-1}\right)=$$
$$(d-1)^{-(k-1)/2}\varrho^k(\underbrace{\beta_{k+1}-\frac{\lambda}{\sqrt{d-1}}\beta_k+\beta_{k-1}}_{0}+(\varrho-1)\beta_{k+1}+(\varrho^{-1}-1)\beta_{k-1}).$$

Thus

$$\sum_{k\geq 1}d(d-1)^{k-1}\delta_k^2\leq r\sum_{k\geq 1}\varrho^{2k}\left((\varrho-1)\beta_{k+1}+(\varrho^{-1}-1)\beta_{k-1}\right)^2.$$

Using that $\varrho^{-1}-1=(1-\varrho)/\varrho\leq 2(1-\varrho)$ and the fact that β_k is bounded we obtain that

$$\sum_{k\geq 1}d(d-1)^{k-1}\delta_k^2\leq C(1-\varrho)^2\sum_{k\geq 1}\varrho^{2k}=C(1-\varrho)^2\frac{\varrho^2}{1-\varrho^2}=C\frac{\varrho^2}{1+\varrho}(1-\varrho)\leq C(1-\varrho),$$

where C might depend on d and λ, but not on ϱ. Therefore the above sum tends to 0 as $\varrho\to 1-$. Similar calculation shows that $\delta_0\to 0$, too. Therefore $\delta_0^2+\sum_{k\geq 1}d(d-1)^{k-1}\delta_k^2\to 0$. Choosing ϱ sufficiently close to 1 and rescaling α_k complete the proof. □

5.3 Independent sets

We have seen that there exist finitely many real numbers $\alpha_0, \alpha_1, \ldots, \alpha_N$ such that the Gaussian process

$$X_v = \sum_{k=0}^{N} \sum_{u:d(v,u)=k} \alpha_k Z_u \tag{5.10}$$

on T_d is *almost* a Gaussian wave function with eigenvalue $\lambda = -2\sqrt{d-1}$. In this section we present different approaches to produce independent sets on T_d using the random variables X_v. In each case the decision whether a given vertex v is chosen for the independent set will depend (in a measurable and invariant way) only on the values of the random variables X_u, $d(v,u) < N'$, where N' is some fixed constant. Therefore the obtained random independent set will be a factor of the i.i.d. process Z_v. Moreover, whether a given vertex v is chosen will depend only on the values in the $N + N'$-neighborhood of v. It follows that the same random procedure can be carried out on any d-regular finite graph G, and the probability that a given vertex is chosen will be the same provided that the girth of G is sufficiently large. Actually, it is enough to assume that G has "essentially large girth", that is, it has $o(n)$ number of small cycles.

So we can work on the regular tree T_d, whence we can actually replace the process (5.10) with the Gaussian wave function for $-2\sqrt{d-1}$. So from this point on, X_v, $v \in V(T_d)$ will denote the Gaussian wave function for $-2\sqrt{d-1}$. Our method works best when the degree d is equal to 3, so we start with the analysis of that case.

5.3.1 The 3-regular case

Let $d = 3$, then $\lambda = -2\sqrt{2}$ and the covariance sequence of X_v is

$$\sigma_0 = 1; \sigma_1 = \frac{-2\sqrt{2}}{3}; \sigma_2 = \frac{5}{6}; \ldots .$$

First approach. We choose those vertices v for which $X_v > X_u$ for each neighbor $u \in N(v)$.

We need to compute the probability

$$P\left(X_{v_0} > X_{v_1}; X_{v_0} > X_{v_2}; X_{v_0} > X_{v_3}\right),$$

where v_0 is an arbitrary vertex with neighbors v_1, v_2, v_3. We will use the fact that if (Y_1, Y_2, Y_3) is a non-degenerate multivariate Gaussian, then the probability that each Y_i is positive can be expressed in terms of the pairwise correlations as follows:

$$P\left(Y_1 > 0; Y_2 > 0; Y_3 > 0\right) = \frac{1}{2} - \frac{1}{4\pi} \sum_{1 \le i < j \le 3} \arccos\left(\operatorname{corr}(Y_i, Y_j)\right). \tag{5.11}$$

Let $Y_i = X_{v_0} - X_{v_i}$, $i = 1, 2, 3$, then we have

$$\operatorname{corr}(Y_1, Y_2) = \frac{\operatorname{cov}(Y_1, Y_2)}{\sqrt{\operatorname{var}(Y_1)\operatorname{var}(Y_2)}} = \frac{1 + \sigma_2 - 2\sigma_1}{2 - 2\sigma_1} = \frac{11 + 8\sqrt{2}}{12 + 8\sqrt{2}} = \frac{1 + 2\sqrt{2}}{4}.$$

The other two correlations are the same, thus we obtain that

$$P\left(v_0 \text{ is chosen}\right) = \frac{1}{2} - \frac{3}{4\pi} \arccos\left(\frac{1 + 2\sqrt{2}}{4}\right) = 0.4298245...$$

In fact, we can add further vertices to this independent set. Let us call a vertex v *addable* if neither v, nor any of its neighbors are chosen. A good portion of the addable vertices can be actually added to our independent set. Simulation showed that the probability that a vertex is addable is about 0.005. When we looked at the connected components of addable vertices, it turned out that most of the components have size 1. Of course, these isolated addable vertices can all be added to our independent set, and some portion of the other components, too. This pushes up the probability to about 0.434. Other local modifications can be made to the independent set, too, since in some cases we can replace one vertex in the independent set with two of its neighbors. We omit the details here, but simulation suggests that with these improvements we can get a bound as good as 0.436. Computing how much we can really gain with these modifications seems very hard, though.

Second approach. We fix some threshold $\tau \in \mathbb{R}$ and we delete those vertices v for which $X_v < \tau$, then we consider the connected components of the remaining graph. If a component is small (its size is at most some fixed N'), then we choose an independent set of size at least half the size of the component. We can do this in a measurable and invariant way. For example, we partition the component into two independent sets (this partition is unique, since each component is connected and bipartite), if one is larger than the other, we choose the larger, if they have equal size, we choose the one containing the vertex v with the largest X_v in the component. If a component is large, then we simply do not choose any vertex from that component. (The idea is to set the parameter τ in such a way that the probability of large components is very small.)

We used computer to simulate the procedure described above. Setting $\tau = -0.12$ and $N' = 200$ the simulation showed that the probability that a given vertex is chosen is above 0.438. In what follows we will show how one can estimate this probability.

From this point on, we will assume that τ is above the critical threshold, that is, each component is finite almost surely. It follows that with probability arbitrarily close to 1 the component of any given vertex has size at most N' provided that N' is sufficiently large. Let p_s denote the probability that the component of a given vertex has size s. (If a vertex is deleted, then we say that its component has size 0. Thus p_0 is simply the probability that $X_v < \tau$.) If a component has size $2k-1$ for some $k \geq 1$, then we choose at least k vertices from the component. If a component contains an even number of vertices, then we choose at least half of the vertices. Thus the probability that a vertex is chosen (in the limit as $N' \to \infty$) is at least

$$\sum_{k=1}^{\infty} \frac{k}{2k-1} p_{2k-1} + \frac{1}{2}\left(1 - p_0 - \sum_{k=1}^{\infty} p_{2k-1}\right) = \frac{1}{2}(1-p_0) + \sum_{k=1}^{\infty} \frac{1}{2(2k-1)} p_{2k-1}. \tag{5.12}$$

If $\tau = 0$, then $p_0 = 1/2$. We can even compute the exact value of p_1. We notice that $X_{v_1} < 0$, $X_{v_2} < 0$ and $X_{v_3} < 0$ imply that $X_{v_0} \geq 0$, because we have a Gaussian wave function with negative eigenvalue. Thus using (5.11) we obtain

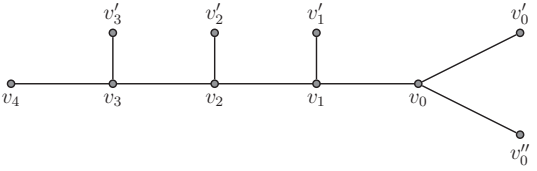
$$p_1 = P\left(X_{v_0} \geq 0; X_{v_1} < 0; X_{v_2} < 0; X_{v_3} < 0\right) = P\left(X_{v_1} < 0; X_{v_2} < 0; X_{v_3} < 0\right) = \\ \frac{1}{2} - \frac{3}{4\pi} \arccos\left(\operatorname{corr}(X_{v_1}, X_{v_2})\right) = \frac{1}{2} - \frac{3}{4\pi} \arccos\left(\frac{5}{6}\right).$$

Using this and the trivial estimates $p_{2k-1} > 0$ for $k \geq 2$, (5.12) yields the following lower bound:

$$\frac{1}{2} - \frac{3}{8\pi} \arccos\left(\frac{5}{6}\right) = 0.4300889...$$

As far as we know, this is the best bound the proof of which needs no computer assistance whatsoever.

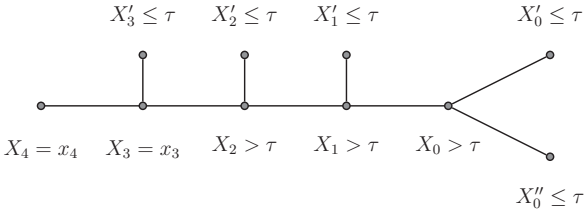
Now we present the best bound we could obtain with a rigorous (but computer assisted) proof. Let τ now be some fixed negative number. Let $k \geq 0$ be an integer and let us fix a path of length $k+1$ in T_3: $v_0, v_1, \ldots, v_{k+1}$. For $1 \leq i \leq k$ let the neighbor of v_i different from v_{i-1} and v_{i+1} be v_i'. Finally, the two neighbors of v_0 different from v_1 are v_0' and v_0''. The random variables assigned to v_i, v_i' and v_i'' will be denoted by X_i, X_i' and X_i'', respectively.



The function $f_k : \mathbb{R}^2 \to [0,1]$ is defined as the following conditional probability:

$$f_k(x_{k+1}, x_k) = P\left(X_i > \tau, 0 \leq i \leq k-1; X_i' \leq \tau, 0 \leq i \leq k; X_0'' \leq \tau | X_{k+1} = x_{k+1}; X_k = x_k\right).$$

The figure below shows the case $k = 3$.



There is a recursive integral formula for these functions. Remark 5.8 says that there exists a standard Gaussian Z_k independent from X_{k+1}, X_k such that

$$X_{k-1} = -\sqrt{2}X_k - \frac{1}{2}X_{k+1} - \frac{1}{2\sqrt{3}}Z_k \quad \text{and}$$
$$X_k' = -\sqrt{2}X_k - \frac{1}{2}X_{k+1} + \frac{1}{2\sqrt{3}}Z_k.$$

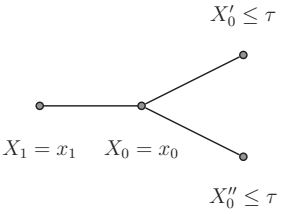
This yields the following formula for the conditional probability $f_k(x_{k+1}, x_k)$ for $k \geq 1$:

$$f_k(x_{k+1}, x_k) = \int_{-\infty}^{-|2\sqrt{6}x_k + \sqrt{3}x_{k+1} + 2\sqrt{3}\tau|} \phi(z_k) f_{k-1}\left(x_k, -\sqrt{2}x_k - \frac{1}{2}x_{k+1} - \frac{1}{2\sqrt{3}}z_k\right) \mathrm{d}z_k,$$

where $\phi(t) = e^{-t^2/2}/\sqrt{2\pi}$ is the density function of the standard normal distribution. As for the case $k = 0$, we have

$$f_0(x_1, x_0) = g_0(2\sqrt{2}x_0 + x_1) = \int_{-(2\sqrt{6}x_0 + \sqrt{3}x_1 + 2\sqrt{3}\tau)}^{2\sqrt{6}x_0 + \sqrt{3}x_1 + 2\sqrt{3}\tau} \phi(z_0)\, \mathrm{d}z_0.$$

(If $a > b$, then $\int_a^b$ is 0.) The case $k = 0$:



Our goal is to compute the probability that all the numbers on a given path in T_3 are above τ and all the numbers on the adjacent vertices are below τ. Let us denote this probability by p'_s when the given path contains s vertices (that is, it has length $s-1$). On the one hand, we have

$$p_1 = p'_1;\ p_3 = 9p'_3;\ \ p_{2k-1} \ge (2k-1)\cdot 3\cdot 4^{k-2}p'_{2k-1},\ k \ge 2.$$

On the other hand, if $s \ge 2$, then for any integer $0 \le m \le s-2$

$$p'_s = \int_\tau^\infty \int_\tau^\infty \phi_2(u,v) f_m(u,v) f_{s-2-m}(v,u)\,\mathrm{d}v\,\mathrm{d}u,$$

where ϕ_2 is the density function of the 2-dimensional centered normal distribution with covariance matrix $\begin{pmatrix} 1 & \sigma_1 \\ \sigma_1 & 1 \end{pmatrix}$ with $\sigma_1 = -2\sqrt{2}/3$. As for $s = 1$,

$$p_1 = p'_1 = \int_{-\infty}^\tau \int_\tau^\infty \phi_2(u,v) f_0(u,v)\,\mathrm{d}v\,\mathrm{d}u.$$

Now let $\tau = -0.086$, then $p_0 = 0.465733....$ According to Theorem 5.10, the components are almost surely finite for this τ. Using computer we obtained the following estimates for p'_1, p'_3, p'_5:

$$p'_1 \ge 0.327277;\ p'_3 \ge 0.002558;\ p'_5 \ge 0.000264024$$

$$p'_3 = \int_\tau^\infty \int_\tau^\infty \phi_2(u,v) f_0(u,v) f_1(v,u)\,\mathrm{d}v\,\mathrm{d}u,$$
$$p'_5 = \int_\tau^\infty \int_\tau^\infty \phi_2(u,v) f_1(u,v) f_2(v,u)\,\mathrm{d}v\,\mathrm{d}u.$$

It follows that

$$p_1 \ge 0.327277, p_3 \qquad\qquad \ge 0.023022, p_5 \ge 0.0158414.$$

These estimates and (5.12) yield the following bound: 0.4361.

5.3.2 The $d \ge 4$ case

The methods presented above for finding independent sets in T_3 work for regular trees with higher degree, too. However, estimating the obtained bounds seems to be a very hard task. According to our computer simulation the second approach with $\tau = -0.04$ yields a lower bound 0.3905 for $d = 4$, where the current best bound is 0.3901.

When the degree is higher than 4, our approach is not as efficient as previous approaches in the literature.

Chapter 6

Invariant random perfect matchings in Cayley graphs

We prove that every non-amenable Cayley graph admits a factor of IID perfect matching. We also show that any connected d-regular vertex transitive infinite graph admits a perfect matching. The two results together imply that every Cayley graph admits an invariant random perfect matching.

A key step in the proof is a result on graphings that also applies to finite graphs. The finite version says that for any partial matching of a finite regular graph that is a good expander, one can always find an augmenting path whose length is poly-logarithmic in one over the ratio of unmatched vertices.

Let Γ be a finitely generated group, and G a locally finite Cayley graph of Γ. An invariant random subgraph on G is a probability distribution on the set of subgraphs of G that is invariant under the natural action of Γ on G.

A factor of IID is a particular way of defining an invariant random subgraph. We only sketch the definition here. First each vertex gets a random number in $[0, 1]$, independently and uniformly. Then each vertex makes a deterministic decision on how the subgraph looks like in its neighborhood, based on what it sees from itself as center. Since each vertex uses the same rule, the distribution of the resulting subgraph is automatically invariant under the action of Γ.

Instead of subgraphs, one can also define vertex colorings, or more general structures on G. The general name for such a random process is a factor of IID process. Invariant random processes, and in particular factor of IIDs on Cayley graphs have received considerable attention recently. Standard percolations are trivially factor of IID processes, as well as the free and the wired minimal spanning forests. Another example is the recent solution of the measurable von Neumann problem by Gaboriau and Lyons (see [26]). They show that every non-amenable Cayley graph admits a factor of IID 4-regular tree.

It is a long standing open problem to determine the maximum density $i(G)$ of a factor of IID independent subset of a regular tree (mentioned e.g. on the webpage of David Aldous [1]). The exact value is unknown, though it is known to be less than 0.46. Note that trees are bipartite and thus have independent sets of density 1/2, but the resulting process can not be a factor of IID. The related open question is to determine the limit of the ratio $i(G(n, d))$ of the largest independent subset in n vertex d-regular random finite graphs, as n goes to infinity. Bayati, Gamarnik, and Tetali in [4] have shown that the limit exists, and the above mentioned modeling phenomenon shows that its value is at least $i(\mathcal{T}_d)$ where $\mathcal{T}_d$ is the d-regular infinite

[1] `http://www.stat.berkeley.edu/~aldous/Research/OP/inv-tree.html`

tree. A conjecture of Balazs Szegedy (see Conjecture 7.13 in [31]) claims that this limit is in fact equal to $i(\mathcal{T}_d)$.

In this paper we settle the analogous question for the maximum density of independent edge sets in non-amenable Cayley graphs. An independent edge set in a graph is usually referred to as a matching. An obvious upper bound on the density of a matching is that of the perfect matching, i.e. where every vertex is covered by an edge. We show that in our case one can actually achieve the maximum possible density, that is, one can construct a perfect matching as a factor of IID.

Theorem 6.1. *Let Γ be a finitely generated non-amenable group with finite symmetric generating set S. Let $G = Cay(\Gamma, S)$ denote the associated Cayley graph. Then there is a factor of IID on G that is almost surely a perfect matching.*

This extends the result of Lyons and Nazarov [47] who proved the same statement for bipartite non-amenable Cayley graphs.

In particular, every non-amenable Cayley graph admits an invariant random perfect matching. Jointly with Abért and Terpai, the authors showed the following theorem.

Theorem 6.2. *Every infinite vertex transitive graph G has a perfect matching.*

Abért and Terpai kindly suggested to include the result in this paper. Now, following an observation of Conley, Kechris, and Tucker-Drob ([15]) this implies that every amenable Cayley graph admits an invariant random perfect matching. Thus, together with Theorem 6.1 we get the following.

Corollary 6.3. *Every Cayley graph admits an invariant random perfect matching.*

The basic strategy of the proof of Theorem 6.1 is similar to what Lyons and Nazarov use to prove the bipartite case, and what has been used by Elek and Lippner [24] to construct almost maximal matchings. We define a sequence of partial matchings, each of which is obtained from the previous one by flipping a sequence of augmenting paths. To show that this sequence "converges" to a limit perfect matching, one has to show that edges do not change roles too often. The crucial step is to bound the length of the shortest augmenting path in terms of the ratio of unmatched vertices.

Our main contribution is establishing this bound for general graphs. When applying the result on finite graphs, we get the following theorem, that is of independent interest in computer science.

Theorem 6.4. *For any $c_0 > 0$ and $d \geq 3$ integer, there is a constant $c = c(c_0, d)$ that satisfies the following statement. If a partial matching in a c_0-expander d-regular graph leaves at least ε ratio of all vertices unmatched, there is an augmenting path of length at most $c \log^3(1/\varepsilon)$, or there is a set of vertices $H \subset G$ such that $|H| \geq 3$, $|H|$ is odd, and the number of edges leaving H is at most d.*

Remark 6.5. The theorem remains true even if there are only two unmatched vertices. This may be surprising at first, but in fact the condition that any odd set H has at least d edges leaving it easily implies the conditions of Tutte's theorem, so such graphs always have perfect matchings.

In the bipartite case, such a bound has actually already been observed in [38] by Jerrum and Vazirani, who used it to give a sub-exponential approximation scheme for the permanent. They remark in the same paper that a similar bound for general graphs would be desirable, as

it would lead to an approximation scheme for the number of perfect matchings for arbitrary graphs. In a subsequent paper we shall work out the details of this application, together with a generalization of Theorem 6.4 to non-regular graphs.

The outline of the paper is as follows. In Section 6.2 we show the existence of perfect matchings in vertex transitive graphs. In Section 6.3 we prove that in a non-amenable Cayley graph there is a factor of IID that is a perfect matching, modulo a variant of Theorem 6.4, whose proof we postpone to Section 6.4.

6.1 Notation and definitions

Let G be a simple graph, either finite or infinite. The vertex and edge set of G will be denoted by $V(G)$ and $E(G)$ respectively.

Definition 6.6. A *matching* in G is a subset $M \subset E(G)$ such that any vertex x is adjacent to at most one edge $e \in M$. We will denote by $V(M)$ the set of vertices that are matched, i.e. that are adjacent to an edge in M. A matching is *perfect* if $V(M) = V(G)$.

Definition 6.7. Given a graph G with a matching M, an *alternating path* is a path $x_0x_1 \dots x_k$ in G such that every second edge belongs to M. An alternating path is called an *augmenting* path if its first and last vertices are not matched.

If $x, y \in V(G)$ are unmatched vertices and p is an augmenting path connecting x and y, then we can define a new matching $M' = M(p) = M \circ E(p)$ as the symmetric difference of the old matching M and the set of edges of p. The new matching will then satisfy $V(M') = V(M) \cup \{x, y\}$.

Let (X, μ) be a standard Borel probability measure space with a non-atomic probability measure μ.

Definition 6.8. A *graphing* on X is a graph $\mathcal{G}$ such that $V(\mathcal{G}) = X$, and where $\mathcal{G}(E) \subset X \times X$ is a symmetric measurable subset, such that if $A, B \subset X$ are measurable subsets and $f : A \to B$ a measurable bijection whose graph $\{(x, f(x)) : x \in A\}$ is a subset of $E(\mathcal{G})$, then $\mu(A) = \mu(B)$.

There is a natural way to measure the size of edge sets in a graphing. If an edge set is given by a measurable bijection $f : A \to B$ as before, then the size of this edge set is defined to be $\mu(A)$. This extends to a measure on all measurable edge sets. In particular this implies that if $\mathcal{H}$ is a sub graphing of $\mathcal{G}$ then the size of the edge set of $\mathcal{H}$ can be computed by the formula

$$|E(\mathcal{H})| = \frac{1}{2} \int_X \deg_{\mathcal{H}}(x) d\mu(x). \tag{6.1}$$

A measurable matching (or matching for short) in $\mathcal{G}$ is a measurable subset $M \subset E(\mathcal{G})$ such that every vertex is adjacent to at most one edge in M. A matching is *almost everywhere perfect* if $\mu(V(\mathcal{G}) \setminus V(M)) = 0$. In this paper we will only be interested in almost everywhere perfect matchings, and will refer to them as perfect matchings for short.

A graphing $\mathcal{G}$ is a c_0-expander if for every measurable set $H \subset V(\mathcal{G})$ we have $|E(H, V(\mathcal{G}) \setminus H)| \geq c_0|H||V(\mathcal{G}) \setminus H|$, where $E(A, B)$ denotes the set of edges having one endpoint in A and one endpoint in B.

Let Γ be a finitely generated group, and $S \subset \Gamma$ a finite symmetric generating set, and $G = Cay(\Gamma, S)$ the associated Cayley graph, that is $g \in \Gamma$ is connected to gs for every $s \in S$. Γ acts on itself by left multiplication, and this naturally extends to a left action on $X =$

$[0,1]^{V(G)} = [0,1]^\Gamma$ by $gx(\gamma) = x(g^{-1}\gamma)$. The latter action is called the Bernoulli shift of Γ. We can equip X with a probability measure μ which is the product of the Lebesgue measure in each coordinate. It is easy to see that the Bernoulli shift action is measure preserving.

Γ also naturally acts from the left on $Y = \{0,1\}^{E(G)}$ whose elements can be considered as subsets of $E(G)$. We can also equip Y with the product of uniform measures on the coordinates.

Definition 6.9. In our context a *factor of IID* is a measurable, Γ equivariant map $\phi : X \to Y$.

Definition 6.10. The graphing $\mathcal{G}$ associated to the Bernoulli shift and S is given by $\mathcal{G}(V) = X$ and $\mathcal{G}(E) = \{(x,y) \in X \times X : \exists s \in S, s^{-1}(x) = y\}$. The connected component of almost any point $x \in X$ is isomorphic to the Cayley graph G.

Claim 6.11. *There is a one-to-one correspondence between measurable subsets $F \subset E(\mathcal{G})$ and factor of IIDs $\phi : X \to Y$.*

Proof. Let $F \subset E(\mathcal{G})$ be a measurable subset and $f : E(\mathcal{G}) \to \{0,1\}$ its characteristic function. Define $\phi_F : X \to Y$ by the following formula.

$$\phi_F(x)(g, gs) = f(s^{-1}g^{-1}x, g^{-1}x).$$

Then

$$(h\phi_F(x))(g, gs) = \phi_F(x)(h^{-1}g, h^{-1}gs) = f(s^{-1}g^{-1}hx, g^{-1}hx) = \phi_F(hx)(g, gs),$$

so we do get a factor of IID.

Conversely, given a factor of IID ϕ, one can define a subset $F_\phi \subset E(\mathcal{G})$ by choosing the edge $s^{-1}x, x$ to be part of F_ϕ if and only if $\phi(x)(id, s) = 1$. □

Remark 6.12.

- From this construction it is clear that F is an almost everywhere perfect matching if and only if ϕ is a factor of IID perfect matching.
- There is an entirely analogous correspondence between measurable subsets of $V(\mathcal{G})$ and factor of IIDs $\phi : X \to \{0,1\}^\Gamma$. In Lemma 2.3 of [47] then translates into the fact that if the Cayley graph G is non-amenable then there is a $c_0 > 0$ depending only on the expansion of G, such that the graphing $\mathcal{G}$ associated to the Bernoulli shift is a c_0-expander.

6.2 Perfect matchings in vertex transitive graphs

Let $G(V, E)$ be an infinite, connected, d-regular, vertex transitive graph. In this section we show that G has a perfect matching. The proof is done in three steps.

Definition 6.13. A *cut* is a partition of V into a nonempty finite set A and its complement $A^c = V \setminus A$. The size of the cut is the number of edges between A and its complement. A *best cut* is a cut with minimum size.

Lemma 6.14. *Suppose $A, B \subset V$ are different finite subsets defining best cuts. Then each of the sets $A \setminus B$, $B \setminus A$, $A \cup B$, and $A \cap B$ is either empty or defines a best cut.*

Proof. Let $X = A \setminus B, Y = B \setminus A, Z = A \cap B, W = V \setminus (A \cup B)$. Then

$$\begin{aligned}|E(X, X^c)| + |E(Y, Y^c)| = \\ = 2|E(X, Y)| + |E(X, Z)| + |E(X, W)| + |E(Y, Z)| + |E(Y, W)| \leq \\ \leq 2|E(X, Y)| + |E(X, Z)| + |E(X, W)| + |E(Y, Z)| + |E(Y, W)| + 2|E(Z, W)| = \\ = |E(X \cup Z, Y \cup W)| + |E(Y \cup Z, X \cup W)| = |E(A, A^c)| + |E(B, B^c)|\end{aligned}$$

This shows that the cuts defined by X and Y are together at most twice the size of the best cut, hence they must be best cuts as well. (Or empty sets.) A similar argument works for Z and W (or rather $X \cup Y \cup Z$, since that is the finite set) as well. □

Lemma 6.15. *The size of the best cut in G is d.*

Proof. Let X be a smallest finite set that defines a best cut. For any pair of vertices $x, y \in X$ there is an automorphism of G that takes x to y. Let Y be the image of X under this automorphism. Then clearly Y also defines a best cut, hence $X \setminus Y$ is also a best cut. But $|X \setminus Y| < |X|$ contradicting the minimality of X, unless $X = Y$. Hence the graph spanned by X is vertex transitive. If $|X| < d$, then the number of edges leaving X is at least $|X|(d-|X|+1) \geq d$ and we are done. If $|X| \geq d$, then since G is connected, there is an edge between a vertex $x \in X$ and $V \setminus X$. But then by vertex transitivity of X, there is such an edge from every single vertex of X, giving the desired lower bound $|E(X, X^c)| \geq |X| \geq d$. □

Corollary 6.16. *Since the number of edges leaving any finite set Y is at most $d|Y|$, and the number of edges entering any finite set X is at least d, we get that the number of finite of components of $G \setminus Y$ is at most $|Y|$.*

Now we are ready to show the existence of perfect matchings in infinite vertex transitive graphs.

Proof of Theorem 6.2. By compactness it is sufficient to show that any finite subset $X \subset V$ can be covered by a matching in G. So assume for contradiction that there is no matching in G that covers a given finite set X.

Let us construct an auxiliary finite graph $G'(V', E')$ as follows. Let $V' = X \cup \partial X \cup M$ where ∂X is the outer vertex boundary of X and M is a non-empty set of new vertices such that $|V'|$ is even. We define the edge set E' to contain all original edges spanned by $X \cup \partial X$, furthermore we add all edges in $\partial X \cup M$ to make it a clique.

If G' has a perfect matching, then just keeping those edges of the matching that intersect X gives a matching in G that covers X. So we can assume that G' does not have a perfect matching. Then by Tutte's theorem there is a set $Y \subset V'$ such that the number of odd components of $G' \setminus Y$ is greater than $|Y|$. But since $|V'|$ is even, we actually get that the number of odd components of $G' \setminus Y$ is at least $|Y| + 2$.

The vertices of $\partial X \cup M$ are always in a single component. Thus we can assume that Y is disjoint from M, since removing vertices of M from Y affects at most one component while reducing the size of Y. Then Y can be thought of a subset of V, and it is easy to see that any finite component of $G' \setminus Y$ is also a finite component of $G \setminus Y$, except perhaps for the one component containing M. Still, this means that $G \setminus Y$ has at least $|Y| + 1$ odd components, all of which are finite, contradicting the previous Corollary. □

Later we will need a slight strengthening of Lemma 6.15. We say that a *real cut* is a cut where the finite set has at least 2 elements.

Lemma 6.17. *The size of the smallest real cut is bigger than d, unless every vertex of G is in a unique d-clique.*

Proof. Suppose the size of the smallest real cut is d, and let X be a smallest finite set that defines a smallest real cut. It is clear that $|X| > 2$ since a set of size 2 defines a cut of size at least $2d - 2 > d$. As before, let $x, y \in X$ and let Y be the image of X under an automorphism taking x to y. We are going to distinguish between three cases according to the size of $X \setminus Y$, which is the same as the size of $Y \setminus X$.

If they have more than 1 element each, then they also real cuts and hence by Lemma 6.15 they are also smallest real cuts, contradicting the minimality of X.

If they are both of size 1, then $|X \cap Y|$ and $|X \cup Y|$ both have to be bigger than 1, hence they are also smallest real cuts, again contradicting the minimality of X.

Thus $|X \setminus Y| = 0$, hence $X = Y$, so just like in the proof of Lemma 6.15 we get that X itself is vertex transitive. Thus, by connectivity, each vertex of X has an edge leaving X. Thus if $|X| \geq d+1$ then we are done. So $|X| \leq d$ and thus the number of edges leaving X is at least $|X|(d - |X| + 1)$. This is strictly greater than d, unless $|X| = 1$ or X is a clique of size d. The first is clearly not the case since X is a real cut. Thus X is a d-clique. Then, of course, by transitivity every vertex of G is in a d-clique.

Finally, it is not possible that a vertex is contained in more than one d-clique. If two different d-cliques A and B intersect then by degree of the vertices in the intersection we see that $|A \cap B| = d - 1$. Let $\{a\} = A \setminus B$ and $\{b\} = B \setminus A$. If a and b would be neighbors then the graph would not be connected. Thus a has to have one neighbor c outside of B. But c cannot be connected to vertices in $A \cap B$, so A is the only d-clique that contains a. But by transitivity each vertex has to be contained in the same number of d-cliques, contradicting our setup. Thus two different d-cliques cannot intersect. □

Corollary 6.18. *If the size of the smallest real cut in G is exactly d then there is a perfect matching in G that is invariant under the automorphism group of G. This matching is given by choosing the unique edge from each vertex that leaves the d-clique the vertex is contained in.*

6.3 Factor of iid perfect matchings via Borel graphs – the proof of Theorem 6.1

Let Γ be a finitely generated non-amenable group, S a finite symmetric generating set of size $|S| = d$, and G the associated Cayley graph. We want to construct a factor of IID perfect matching in G.

If the size of the smallest real cut in G is equal to d, then by Corollary 6.18 there is a fixed perfect matching in G that is invariant under the action of the automorphism group, and each vertex can decide which edge to choose by observing its own 1-neighborhood, so this is clearly a factor of IID matching and we are done.

Thus we can assume that the smallest real cut in G is at least of size $d + 1$. Let $\mathcal{G}$ be the graphing associated to the Bernoulli shift, as in Definition 6.10. By Claim 6.11 and Remark 6.12 it follows that $\mathcal{G}$ is a c_0-expander for some $c_0 > 0$ depending only on G. Hence $\mathcal{G}$ is admissible in the sense of Definition 6.20.

By Remark 6.12 it is now sufficient to prove that $\mathcal{G}$ has an almost everywhere perfect matching. Proposition 1.1 in [24] shows that there exists a sequence of matchings $M_0, M_1, M_2, \cdots \subset \mathcal{G}$ such that a) there are no augmenting paths of length $2k + 1$ in M_k and b) each M_k is obtained from M_{k-1} by a sequence of flipping augmenting paths of length at most $2k + 1$. We would

like to construct an almost everywhere perfect matching as a limit of the M_ks. In order to do this, we have to show that, except for a measure zero set, the status of any edge changes only finitely many times during the process, so we can take a "pointwise" limit of the sequence to obtain a matching that covers but a zero measure subset of X.

Let us denote by U_k the set of unmatched vertices in M_k. Then in the process of getting M_{k+1} from M_k we are flipping augmenting paths starting and ending in U_k. Furthermore each vertex of U_k can be only used once as an endpoint of an augmenting path, since after that it becomes a matched vertex. Any edge that changes status between M_k and M_{k+1} has to be part of an augmenting path at least once. Thus the total measure of status changing edges in this step is at most $(2k+3)|U_k|$. If we can show that $\sum_k (2k+3)|U_k| < \infty$ then by the Borel-Cantelli lemma the measure of edges that change status infinitely many times is zero, and we are done.

We have seen that $\mathcal{G}$ is admissible. Let $\varepsilon = |U_k|$. Then by Theorem 6.21 there is a constant $c = c(c_0, d)$ depending only on the expansion of $\mathcal{G}$ and the degree d, such that there is an augmenting path of length at most $c \log^3(1/\varepsilon)$ in M_k. But by definition we know that this has to be longer than $2k+1$. Thus we get $2k+1 \leq c \log^3(1/\varepsilon)$ or equivalently

$$|U_k| = \varepsilon < \exp\left(-\left(\frac{2k+1}{c}\right)^{1/3}\right).$$

This is clearly small enough to guarantee that $\sum_k (2k+3)|U_k| < \infty$ and thus complete the proof of Theorem 6.1. □

Corollary 6.19. *Every d-regular infinite Cayley graph has an invariant random perfect matching.*

Proof. For amenable graphs Conley, Kechris and Tucker-Drob observed in Proposition 7.5 of [15] that Theorem 6.2 implies the existence of invariant random matchings.

Since a factor of IID perfect matching is automatically an invariant random perfect matching, Theorem 6.1 completes the non-amenable case. □

6.4 Short alternating paths in expanders

Let $G(X, E)$ be a d-regular graphing, or a connected, d-regular graph that can either be finite or infinite. We are going to treat these three cases at the same time. When it is necessary to point out differences, we will refer to them as the measurable/finite/countable case respectively.

Definition 6.20. We say that G is *admissible* if it is a c_0-expander, and the smallest real cut into odd sets has size at least $d+1$ (in the sense of Lemma 6.17).

The following theorem includes the statement of Theorem 6.4 and the variant about graphings that is needed for the proof of Theorem 6.1.

Theorem 6.21. *For any $c_0 > 0$ and $d \geq 3$ integer, there is a constant $c = c(c_0, d)$ that satisfies the following statement. Given any admissible measurable (or large finite) graph, and a partial matching with at least ε measure (or fraction) of unmatched vertices, there is an augmenting path of length at most $c \log^3(1/\varepsilon)$.*

Though our main goal is to prove theorems about measurable graphs and finite graphs, we are going to need auxiliary results about infinite, connected d-regular graphs as well. Since the three cases can be handled the same way, we are going to present the proofs at the same time,

pointing out differences when necessary. In the measurable case, everything will be assumed to be measurable, unless explicitly stated otherwise. If $A, B \subset X$ then $E(A, B)$ will denote the set of edges that have one endpoint in A and the other in B. In the measurable case the measure of the set A will be denoted by $|A|$. In the finite case $|A|$ is going to denote the size of A divided by the total number of vertices in the graph. So in both of these cases $0 \leq |A| \leq 1$. In fact, a finite graph can be considered as a graphing with an atomic probability measure. However in the countable case $|A|$ is going to simply denote the size of A. Similarly with edge sets, in the finite and the measurable cases $|E(A, B)|$ will denote the measure of the edge set as defined by the integral (6.1) in Definition 6.8. In the countable case $|E(A, B)|$ will just denote the size of the set $E(A, B)$. If we really want to talk about the actual size of sets in the finite case, we will denote it by $||A||$ and $||E(A, B)||$ respectively.

Let $M \subset E$ be a matching. Then $V(M) \subset X$ shall denote the set of matched vertices. Let $S \subset X \setminus V(M)$ denote a fixed subset of the unmatched vertices and let $F = X \setminus (V(M) \cup S)$ denote the remaining unmatched vertices. We are going to construct alternating paths starting from S in the hope of finding an alternating path connecting two unmatched vertices. Such an alternating path is called an *augmenting path*.

6.4.1 Sketch of the proof

First we give an outline, pointing out the main ideas without introducing the technical definitions. We encourage the reader to read the whole outline before reading the proof, and also to refer back to it whenever necessary. Without understanding the basic outline, many technical definitions will likely be rather unmotivated.

1. We start from a set of unmatched vertices S. Assuming there are no short augmenting paths, we would like to show that the set of vertices (X_n) accessible via n-step or shorter alternating paths grows rapidly, eventually exceeding the size of the whole graph, leading to a contradiction.

2. It will be necessary to keep track of matched vertices accessible via odd paths (head vertices), even paths (tail vertices), or both. In notation $X_n = S \cup H_n \cup T_n \cup B_n$.

3. If there are plenty of edges leaving X_n from T_n or B_n, then the other ends of these edges will be part of X_{n+1}, fueling the desired growth. The first observation is that if this is not the case, then there has to be many tail-tail or tail-both edges.

4. A tail vertex that has another tail- or both-type neighbor will normally become a both-type vertex in the next step. In this case even though X_n does not grow, the set B_n grows within X_n, still maintaining the desired expansion that eventually leads to a contradiction.

5. The problem is that certain tail-vertices will not become both-type even though they possess a both-type neighbor. These will be called the *tough* vertices. The bulk of the proof is about bounding the number of tough vertices. The key idea here is that we can associate to each tough vertex x a distinct subset of B_n called the *family* of x. Families associated to different vertices are pairwise disjoint. (This is done in Section 6.4.3.)

6. There can not be too many tough vertices with large families. On the other hand if a vertex stays tough for an extended amount of time, its family has to grow. These two observation together should be enough to bound the number of tough vertices.

7. The proof proceeds in two rounds from this point. First, if X_n is smaller than half of the graph, then already families larger than $4d(d+1)/c_0$ are too large, and indeed vertices can't be tough too long before they reach this critical family size. Then all the previous observations are valid and X_n grows exponentially as desired. (This is the contents of Theorem 6.24 and the proof is done in Section 6.4.4.)

8. In the second round, when X_n is already quite big, this unfortunately does not work anymore. The bound after which families can be deemed too large grows as $|X \setminus X_n|$ shrinks, and thus vertices can be tough longer and longer before their families become big enough. At this point it becomes necessary to show that the families of tough vertices also grow exponentially fast.

9. In Section 6.4.5 we demonstrate that the dynamics of how a family grows is almost identical to how the sets X_n are growing. In fact families are more or less what can be reached from the tough vertex by an alternating path. But a family lives within an infinite countable graph, hence it is never bigger than "half of the graph", so only the first round is needed to show exponential growth. Hence Theorem 6.24 has a double gain. It proves the first round for X_n, but at the same time it is used to prove fast family growth in the second round.

10. Once we have established exponential family growth, an approach very similar to the proof of the first round is used to complete Theorem 6.21 in Section 6.4.6. The proofs of both rounds employ a method of defining an invariant whose growth is controlled. But the hidden motivation behind the invariant is what we have outlined in this sketch: if X_n doesn't grow then B_n grows. If B_n doesn't grow either then there have to be many tough vertices. If there are many tough vertices then they have to be tough for a long time. But then their families have to become too big. Finally there is no space for all these big families.

11. Unfortunately there is a final twist. When analyzing family growth in Section 6.4.5, we have to introduce certain forbidden edges in each step, through which alternating paths are not allowed to pass momentarily. Hence, to be able to use Theorem 6.24 in this more general scenario, we need to state it in a rather awkward way. Instead of saying that X_n is just what can be reached by alternating paths of length at most n, we need to use a recursive definition of X_n taking into account the forbidden edges in each step. But as it is pointed out in Remark 6.23, if one chooses to have no forbidden edges, X_n just becomes what it was in this sketch.

The proof is organized as follows. In Section 6.4.2 we introduce the basic recursive construction of the X_k sets using the notion of forbidden edges. We state the key Theorem 6.24 that on one hand provides the proof of the first round, and on the other hand will be used to show exponential family growth.

Tough vertices and families are introduced in Section 6.4.3 together with proofs of their basic properties. Then Theorem 6.24 and the first round is proved in Section 6.4.4, using the invariant-technique.

In Section 6.4.5 we show how the growth of a family can be modeled using the forbidden edge construction, and prove exponential growth of families. Finally in Section 6.4.6 we finish the proof of the second round, again using the invariant-technique.

6.4.2 Forbidden edges

We are going to use the following terminology. All alternating paths will start with an unmatched edge, but may end with either kind of edges. If $p = (p_0, p_1, \dots, p_l)$ is an alternating path of length $|p| = l$, then the vertices with odd index will be referred to as the "head" vertices of p and the even index vertices (except for p_0) will be called "tail" vertices. p will be called even if l is even, and odd if l is odd. The last vertex will be denoted by $\mathrm{end}(p) = p_l$. When this doesn't cause confusion, we will also use p to denote just the set of vertices of the path.

Definition 6.22. Assume that for every k we are given a subset of "forbidden" edges $E_k \subset E$. Using this as input data, we shall recursively construct a sequence of vertex sets

$$S = X_0 \subset X_1 \subset X_2 \subset \dots.$$

Suppose we have already defined X_k. Then X_{k+1} is defined as follows. Take a matched edge vw outside of X_k. We are going to include these two vertices in X_{k+1} if and only if there is an even alternating path starting in S whose length is at most $2k+2$, whose last two vertices are v and w in some order while all the previous vertices are in X_k, and, most importantly, the edge on which it leaves X_k does not belong to E_k.

Remark 6.23. This definition implies that each X_k consists of matched pairs, and for any vertex $v \in X_k$ there is an alternating path $p \subset X_k$ such that $p_0 \in S$, $|p| \le 2k$, and $\mathrm{end}(p) = v$. If the E_k are all empty, then X_k consists of all vertices accessible from S via an alternating path of length at most $2k$. First we will show that the size of X_k grows fast.

Theorem 6.24. *Suppose that*

1. $|X_n| \le |X \setminus X_n|$,
2. *there are no augmenting paths of length at most* $2n-1$ *starting in* S, *and*
3. $|E_k| \le d|S|$ *for all* $0 \le k < n$,
4. *the number of non-forbidden edges leaving* X_k *is at least* $1/(d+1)$ *portion of all edges leaving* X_k *for all* $k < n$.

Then

$$|X_n| \ge \frac{c_0^2|S|}{16d^2(d+1)^2}\left(1 + \frac{c_0^3}{128d^3(d+1)^3}\right)^n.$$

Note that the first condition is always satisfied in the countable case, since X_n is always finite.

We will need a more refined classification of the vertices in X_n. First of all, let $\mathfrak{A}_o$ denote the set of all odd alternating paths starting from S, and $\mathfrak{A}_e$ the set of all even alternating paths. For every $n \ge 1$ let us define the following subsets of X_n. Let

$$\tilde{H}_n = \{x \in X_n : \exists p \in \mathfrak{A}_o (1 \le |p| \le 2n, p \subset X_n; \mathrm{end}(p) = x)\},$$

$$\tilde{T}_n = \{x \in X_n : \exists p \in \mathfrak{A}_e (2 \le |p| \le 2n, p \subset X_n; \mathrm{end}(p) = x)\},$$

$$H_n = \tilde{H}_n \setminus \tilde{T}_n,$$

$$T_n = \tilde{T}_n \setminus \tilde{H}_n,$$

$$B_n = \tilde{H}_n \cap \tilde{T}_n.$$

It is important that in these definitions we are not insisting that the paths avoid forbidden edges at any time. The forbidden edges only limit the definition of X_n, but then we want to consider all possible alternating paths within the set.

The last three are the set of head vertices, the set of tail vertices, and those that can be both heads or tails. It is clear that S and T_n are disjoint. As long as there are no augmenting paths of length at most $2n-1$, then S is also disjoint from $\tilde{H}_n$, and thus X_n is a disjoint union of S, H_n, B_n, and T_n. It follows from the definition that $B_1 \subset B_2 \subset \ldots$, furthermore M gives a perfect matching between T_n and H_n, and also within B_n. (Note that this implies $|H_n| = |T_n|$.)

The rough idea of why X_n should grow fast is this. By expansion, even in the presence of forbidden edges, there are plenty of edges leaving X_n. Any edge leaving X_n from $\tilde{T}_n$ adds to the size of X_{n+1} directly. Only edges leaving from H_n cause problems. But since H_n and T_n have the same total degree, any surplus of edges leaving H_n have to be compensated by edges within T_n or between B_n and T_n. Such edges will contribute to the growth of B_n within X_n, and thus implicitly to the growth of X_n.

6.4.3 Combinatorics of alternating paths

In this section we will be mainly concerned about how edges within $T_n \cup S$ and between B_n and $T_n \cup S$ contribute to the growth of B_n.

Lemma 6.25. *If $x, y \in T_n \cup S$ and $xy \in E$ then either $x \in B_{n+1}$ or $y \in B_{n+1}$ or there is an augmenting path of length at most $2n+1$.*

Proof. It is sufficient to prove that either x or y would be in $\tilde{H}_{n+1}$. Let p, respectively q be shortest alternating paths that witness x and $y \in T_n$ respectively. We may assume without loss of generality that $|p| \leq |q|$. Then y cannot lie on p, otherwise there would either be a shorter alternating path witnessing $y \in T_n$, or we would have $y \in \tilde{H}_n$ and not in $T_n \cup S$. Hence adding the xy edge to p we obtain an alternating path of length at most $2n+1$ that witnesses that $y \in \tilde{H}_{n+1}$. □

Edges running between $T_n \cup S$ and B_n are more complicated to handle. If $b \in B_n$ and $t \in T_n \cup S$, but all paths witnessing $b \in \tilde{T}_n$ run through t, then we can't simply exhibit $t \in \tilde{H}_{n+1}$ by adding the bt edge to the end of such a path since it would become self-intersecting. The following definition captures this behavior.

Definition 6.26.

- A vertex $x \in T_n \cup S$ is "tough" if it is adjacent to one or more vertices in B_n, but $x \notin \tilde{H}_{n+1}$.
- An edge $xy \in E$ is "tough" if $x \in T_n \cup S, y \in B_n$ and x is a tough vertex.

TT_n will denote the set of vertices that are tough at time n.

We would like to somehow bound the number of tough vertices. In order to do so, we will associate certain subsets of X_n to each tough vertex in a way that subsets belonging to different tough vertices do not intersect. Then we will show that these subsets become large quickly.

Remark 6.27. We think of n as some sort of time variable, and all the sets evolve as n changes. Usually n will denote the "current" moment in this process. In the following definitions of age, descendent, and family, there will be a hidden dependence on n. When talking about the age or the family of a vertex, we always implicitly understand that it is taken at the current moment.

Definition 6.28. The "age" of a vertex $x \in TT_n$ is $a(x) = n - \min\{k : x \in T_k \cup S\}$.

Definition 6.29. Fix a vertex $x \in TT_n$. A set $D \subset X_n$ has the "descendent property" with respect to x if the following is true. For every $y \in D$ there are two alternating paths p and q starting in x and ending in y, such that

- both start with an unmatched edge, but p is odd while q is even,
- $p, q \subset D \cup \{x\}$,
- $|p| + |q| \leq 2a(x) + 1$.

Sets satisfying the descendent property with respect to x are closed under union.

Definition 6.30. The "family" of a vertex $x \in TT_n$ is the largest set $D \subset X_n$ that satisfies the descendent property. In other words it is the union of all sets that satisfy the descendent property. The family of x is denoted by $F_n(x)$.

Claim 6.31. *If $x \in TT_n$ and xy is a tough edge then y is in the family of x. In particular every tough vertex has a nonempty family.*

Proof. Let p be a path that witnesses $y \in \tilde{T}_n$. Now if p appended by the edge yx would be a path then it would witness $x \in \tilde{H}_{n+1}$. Since this is not the case, x has to lie on p. Let D denote the set of vertices p visits after leaving x. For any point $z \in D$ there are two alternating paths from x to z. One is given by p and the other by going from x to y and then walking backwards on p. Suppose $x = p_{2l}$ and $y = p_{2k}$. The total length of these two paths is $2k - 2l + 1$. Since the age of x by Definition 6.28 is at least $k - l$ we see that $2k - 2l + 1 \leq 2a(x) + 1$. Hence the two paths satisfy all conditions of Definition 6.29 so D has the descendent property with respect to x. Hence by Definition 6.30 x has a non-empty family, in particular y is in the family. □

Claim 6.32. *The family of any tough vertex is a subset of B_n.*

Proof. Let $x \in TT_n$ be a tough vertex and let s be a shortest path witnessing $x \in T_n \cup S$. Let us denote $|s| = 2k$. It is enough to show that the family of x is disjoint from s. Indeed, then for any point y in the family we can take the two types of paths p, q as in Definition 6.29 from x to y. By the age requirement in Definition 6.29 we get that $|p| + |q| \leq 2a(x) + 1 = 2n - 2k + 1$. Hence $|s| + |p| + |q| \leq 2n + 1$ and thus $|s| + |p| \leq 2n - 1$ and $|s| + |q| \leq 2n$. Since these paths run within the family which is disjoint from s, we can append s with p and q respectively to get alternating paths witnessing $y \in \tilde{H}_n$ and $y \in \tilde{T}_n$ respectively.

Now suppose the family of x is not disjoint from s. It is clear that any family consists of pairs of matched vertices. Let i be the smallest index such that the pair s_{2i-1}, s_{2i} is in the family. Then from x there is an odd alternating path p to s_{2i} by Definition 6.29 that runs within the family and its length is at most $2a(x) + 1 \leq 2n - 2k + 1$. Since i was the smallest such index, the path p is disjoint from $s_0, s_1, \ldots s_{2i-1}$. Thus by appending $s_0, s_1, \ldots, s_{2i}$ by the reverse of p we get an alternating path from an unmatched point to x ending in an unmatched edge, whose length is at most $2n - 2k + 1 + 2i \leq 2n + 1$. This path witnesses $x \in \tilde{H}_{n+1}$, contradicting the toughness of x. □

Next we will prove that any vertex can belong to at most one family. We start with a simple lemma about concatenating alternating paths.

Lemma 6.33. *Let p be an even alternating path from x to y and q an odd alternating path from y to z. Then there is an odd alternating path from x to either y or z whose length is at most $|p| + |q|$.*

Proof. If the concatenation of p and q is a path, then we are done. Otherwise let i be the smallest index such that $p_i \in q$. Let $p_i = q_j$. Then $p_0, p_1, \ldots, p_i = q_j, q_{j+1}, \ldots, \mathrm{end}(q)$ is a path from x to z and $p_0, p_1, \ldots, p_i = q_j, q_{j-1}, \ldots q_0$ is a path from x to y. Both have length at most $|p| + |q|$, both of them end with non-matched edges and one of them is clearly alternating. □

Claim 6.34. *Two families cannot intersect.*

Proof. Let $x, y \in TT_n$ be two tough vertices. Assume their families F and G do intersect. Let p, q be shortest alternating paths witnessing $x, y \in T_n \cup S$. Let us choose the shortest among all alternating paths from x to $F \cap G$ that runs within F. Let this path be p' and its endpoint $x' \in F \cap G$. Do the same thing with y to get a path q' from y to $y' \in F \cap G$ lying within G. By symmetry we may assume that $|p| + |p'| \leq |q| + |q'|$.

By the choice of p' we see that the only point on p' that is in G is its endpoint x'. From x' there are two paths, s and t, leading to y within G by Definition 6.29 one of which, say s, can be appended to p' to get an alternating path from x to y. This path $p' \cup s$ clearly starts and ends with a non-matching edge.

Now we are in a situation to apply the previous lemma. p leads from p_0 to x and ends with a matching edge, and $p' \cup s$ leads from x to y and starts and ends with non-matching edges. Thus by the lemma, there is an alternating path from p_0 to either x or y which ends with a non-matching edge. The length of this alternating path is at most $|p| + |p'| + |s|$. But by the choice of p', the choice of q', and by the age requirement in Definition 6.29 we have

$$|p| + |p'| + |s| \leq |q| + |q'| + |s| \leq |q| + |t| + |s| \leq |q| + 2a(y) + 1 = 2n + 1.$$

Thus the alternating path we have found from p_0 to x or y has length at most $2n + 1$ so it witnesses $x \in \tilde{H}_{n+1}$ or $y \in \tilde{H}_{n+1}$. But neither is possible since both x and y are tough, which is a contradiction. □

Corollary 6.35. *There is exactly one tough vertex adjacent to any family.*

Proof. Let $x, y \in TT_n$ and $z \in F_n(x)$. Suppose there is an edge between y and z.Then z is in B_n, hence yz is a tough edge, hence z is in the family of y, but then the two families would not be disjoint, which is a contradiction. □

Let $c_1 = c_1(c_0, d)$ be a constant to be determined later.

Claim 6.36. *Suppose $|F_n(x)| < c_1$, $v \in F_n(x)$, and there is an edge vw such that $w \in B_n \backslash F_n(x)$. Then either $w \in F_{n+c_1}(x)$ or $x \in \tilde{H}_{n+c_1}$. In other words, if a vertex remains tough for an extended period of time, then its family consumes its neighbors.*

Proof. We can assume that $x \notin \tilde{H}_{n+c_1}$ since otherwise we are done. Thus x is still tough at the moment $n + c_1$.

First suppose there is a path $p \in \mathfrak{A}_e, |p| \leq 2n$ that ends in w and does not pass through x. Let $w' \in p$ be the first even vertex on the path that is adjacent to some vertex $v' \in F_n(x)$. Then the initial segment of p up until w' has to be disjoint from $F_n(x)$. By definition, in $F_n(x)$ there has to be an alternating path from x to v' that ends in a matched edge. Extending this path through w' and then the initial segment of p, we get an alternating path from S to x. Its length is obviously at most $|p| + c_1$, hence $x \in \tilde{H}_{n+c_1/2}$ and consequently in $\tilde{H}_{n+c_1}$, and this is a contradiction.

That means that any even path from S to w of length at most $2n$ has to pass through x. Let p be the shortest such path. Let v' be the last vertex of p that is in $F_n(x) \cup \{x\}$.

The vertex v' divides p into two segments, p_1 going from S to v' and p_2 from v' to w. Then $|p_2| = |p| - |p_1| \leq 2n - 2\min\{k : x \in T_k \cup S\} = 2a(x)$, and equality can only happen if $x = v'$. We claim that p_2 becomes part of the family at time $n + c_1$. For any vertex $y \in p_2$ we can either go from x to v' in even steps and then continue along p_2, or go from x to v in even steps and continue backwards on p_2 to y. The total length of these two paths is at most $c_1 + |p_2| + 1 + c_1 \leq 2(a(x) + c_1) + 1$. Since at moment $n + c_1$ the age of x is exactly $a(x) + c_1$, the set $F_n(x) \cup p_2$ will satisfy the descendent property, so this whole set, including w, will be part of $F_{n+c_1}(x)$. □

Definition 6.37. We will say that at moment n the family of the vertex $x \in TT_n$ is *expanding* if there is an edge vw such that $v \in F_n(x)$ and $w \in B_n \setminus F_n(x)$. For any $x \in X$, let $e_n(x)$ be the number of moments $m < n$ such that $0 < |F_m(x)| < c_1$ and at moment m the family was expanding.

Corollary 6.38. *For any $x \in X$ we have $e_n(x) \leq c_1^2$ independently of n.*

Proof. By Claim 6.36 we know that the number of moments in which an expanding family has a fixed size $k < c_1$ is at most c_1. This is because after the first such moment, in c_1 time the family either ceases to exist or strictly grows. Thus for each possible size k there are at most c_1 moments of expansion, and thus there are at most c_1^2 such moments in all. □

6.4.4 Invariants of growth

Now we are ready to start the proof of Theorem 6.24. Let

$$I(n) = |X_n| + |B_n| + \frac{1}{2}\int_X e_n(x)dx.$$

or in the infinite connected case

$$I(n) = |X_n| + |B_n| + \frac{1}{2}\sum_{x \in X} e_n(x).$$

Proposition 6.39. *Suppose that*

1. *$|X_n| \leq |X \setminus X_n|$,*
2. *there are no augmenting paths of length at most $2n - 1$ in X_n, and*
3. *the number of forbidden edges is $|E_k| \leq d|S|$ for all $0 \leq k < n$,*
4. *the number of non-forbidden edges leaving X_k is at least $1/(d + 1)$ portion of all edges leaving X_k for all $k < n$.*

then

$$I(n + 1) \geq \left(1 + \frac{c_0^3}{128d^3(d + 1)^3}\right) I(n).$$

Proof. In the following we shall omit the index n from all our notation, except where it would lead to confusion. Let TT denote the set of tough and TM the set of not-tough vertices within $T \cup S$. The tough vertices are further classified according to their families. TB denotes the tough vertices whose families have size at least c_1. For tough vertices with smaller families, TE shall denote the ones that have expanding families and TG denote the rest. So

$$S \cup T = TM \cup TT = TM \cup (TB \cup TE \cup TG).$$

First let's take a tough vertex $x \in TG$ whose family is small and not expanding. Let $|E(x, F(x))| = k$. By the assumption on the size of the smallest real odd cut we know that the number of edges leaving $x \cup F(x)$ (which is a set of odd size!) is at least $d+1$. But only $d-k$ of these are adjacent to x, so at least $k+1$ have to be adjacent to $F(x)$. None of these edges can lead to B because this is a non-expanding family. Also non of these edges can lead to TT by Corollary 6.35. Hence all these edges have to go to H, TM, or the outside world $O = X^c$. This means that

$$|E(F(x), TG)| = |E(F(x), x)| \leq |E(F(x), H \cup TM \cup O)|. \tag{6.2}$$

By Claim 6.31 we see that any edge between TG and B has to run between a vertex in TG and a member of its family. Thus integrating (6.2) over $x \in TG$ and using that families are pairwise disjoint subsets in B we get that

$$|E(B, TG)| \leq |E(B, H \cup TM \cup O)|$$

For any other tough vertex we bound the number of edges between it and B by the trivial bound d. Adding this to the previous equation we get

$$|E(B, TT)| \leq d|TB| + d|TE| + |E(B, H \cup TM \cup O)| \tag{6.3}$$

We know that $|T| = |H|$ because of the matching, so the total degrees of $S \cup T$ is $d|S|$ more than the total degree of H. The edges between $T \cup S$ and H contribute equally to these total degrees. In the worst case there are no internal edges in H. This boils down to the following estimate.

$$\begin{aligned} |E(H, O)| + |E(H, B)| + d|S| \leq \\ \leq 2|E(T \cup S, T \cup S)| + |E(T \cup S, O)| + |E(TM, B)| + |E(TT, B)|. \end{aligned}$$

Combining it with (6.3), and subtracting $|E(H, B)|$ from both sides we get

$$\begin{aligned} |E(H, O)| + d|S| \leq 2|E(T \cup S, T \cup S)| + 2|E(B, TM)| + \\ + |E(B \cup T \cup S, O)| + d|TB| + d|TE|. \end{aligned}$$

Any vertex in TB has a family of size at least c_1, and all these are disjoint by Claim 6.34 and contained in B. Thus we get that $|TB| \leq |B|/c_1$. Using this and adding $|E(B \cup T \cup S, O)|$ to both sides implies

$$\begin{aligned} |E(X_n, O)| + d|S| \leq 2|E(T \cup S, T \cup S) + 2|E(B, TM)| + \\ + 2|E(B \cup T \cup S, O)| + \frac{d}{c_1}|B| + d|TE|. \end{aligned} \tag{6.4}$$

Any vertex in O_n that is adjacent to $B_n \cup T_n \cup S$ along an edge not in the forbidden set E_n is going to be in X_{n+1}, hence

$$|E(B_n \cup T_n \cup S, O_n) \setminus E_n| \leq d(|X_{n+1}| - |X_n|).$$

By definition, any vertex in TM_n that is adjacent to an edge coming from B_n will be part of B_{n+1} or yield an augmenting path. Also, by Lemma 6.25, any edge in $E(S \cup T_n, S \cup T_n)$ has to be adjacent to a point in $|B_{n+1}| \setminus |B_n|$ or yield an augmenting path. This implies that

$$2|E(T, T)| + 2|E(B, TM)| \leq 2d(|B_{n+1}| - |B_n|).$$

By the 3rd assumption of the proposition we have $|E_n| \leq d|S|$. Plugging all this into (6.4) we get

$$\frac{|E(X_n, O_n) \setminus E_n|}{d} \leq 2(|X_{n+1}| - |X_n|) + 2(|B_{n+1}| - |B_n|) + |TE| + \frac{|B_n|}{c_1} \tag{6.5}$$

By Definition 6.37, for any vertex $x \in TE_n$ we get $e_{n+1}(x) = e_n(x) + 1$, and thus

$$\int_X e_{n+1}(x)dx = \int_X e_n(x)dx + |TE|.$$

Hence the right hand side of (6.5) is exactly $2(I(n+1) - I(n)) + |B_n|/c_1$. Furthermore by the 4th and 1st assumptions of the proposition we have

$$|E(X_n, O_n) \setminus E_n| \geq \frac{|E(X_n, O_n)|}{d+1} \geq \frac{c_0|X_n|(1-|X_n|)}{d+1} \geq \frac{c_0}{2d+2}|X_n|$$

in the measurable case and

$$|E(X_n, O_n) \setminus E_n| \geq \frac{|E(X_n, O_n)|}{d+1} \geq \frac{c_0|X_n|}{d+1} \geq \frac{c_0}{2d+2}|X_n|$$

in the connected infinite case. So in either case we get

$$\frac{c_0}{4d(d+1)}|X_n| - \frac{|B_n|}{2c_1} \leq I(n+1) - I(n)$$

Now we can complete the proof of the proposition. First, choose $c_1 = 4d(d+1)/c_0$. Then, since $|B_n| \leq |X_n$ we get

$$\frac{|X_n|}{2c_1} \leq I(n+1) - I(n).$$

On the other hand, we know from Corollary 6.38 that $e_n(x) \leq c_1^2$. Obviously $e_n(x) = 0$ if $x \in O_n$. Thus $\int_X e_n(x)dx \leq c_1^2|X_n|$. Hence

$$I(n) \leq \left(2 + \frac{c_1^2}{2}\right)|X_n| \leq c_1^2|X_n| \leq 2c_1^3(I(n+1) - I(n)).$$

Substituting $c_1 = 4d(d+1)/c_0$ finishes the proof. □

This proposition implies that $I(n)$ grows exponentially fast. But as we have seen, $|X_n|$ can be bounded from below in terms of $I(n)$. This will imply fast growth of $|X_n|$ too.

Proof of Theorem 6.24.

Since $S \subset X_0$, we have $|S| \leq I(0)$. Then again by Corollary 6.38 we have $I(n) \leq c_1^2|X_n|$. So by Proposition 6.39

$$X(n) \geq \frac{I(n)}{c_1^2} \geq \frac{|S|}{c_1^2}\left(1 + \frac{c_0^3}{128d^3(d+1)^3}\right)^n$$

Substituting $c_1 = 4d(d+1)/c_0$ we get the desired result. □

In the measurable case, when X_n becomes large, the method apparently breaks down. The main problem is that expansion guarantees only $c_0|X_n|(1-|X_n|)$ edges between X_n and O_n. When X_n is large, the $1 - |X_n|$ term will be the dominant. It was crucial to choose c_1 so that the $|B_n|/c_1$ terms becomes comparable to the lower bound coming from expansion. But for

large B_n, hence small $1-|X_n|$, this cannot be done with a constant c_1. The smallest c_1 that has a chance to work is roughly on the scale of $1/\varepsilon$. But then the upper bound for $I(n)$ becomes $(1/\varepsilon)^3$ and all of a sudden the time needed for X_n to exceed $1-\varepsilon$ becomes super-linear in $1/\varepsilon$ instead of the desired poly-logarithmic dependence.

This loss of time comes from the part where we argued that any family grows bigger than c_1 in c_1^2 time. This observation was sufficient for a constant c_1, but is clearly insufficient when $c_1 \approx 1/\varepsilon$. In this part we will show that, in fact, families grow much faster than what Claim 6.36 asserts. It turns out that in a sense families grow exponentially, hence it takes much less time than $(1/\varepsilon)^2$ to reach a size of $1/\varepsilon$. This will allow us to "fix" the argument in Section 6.4.4.

6.4.5 Family business

In this section we shall examine in detail the lifecycle of a family. Let us fix a vertex $x \in X$. At some n_0, this x may become an element of T_{n_0}. Then later it may start to have neighbors in B_{n_1} (for a larger value $n_1 \geq n_0$). At this point it can become tough and start to have a family. This family grows in time, until at some even larger value of n the vertex finally becomes part of B_n. We want to understand the part when x becomes tough and its family starts growing.

To this end we shall recursively define a sequence of "special moments"

$$n_0 \leq n_1 < n_2 < n_3 \dots$$

and an increasing sequence of sets

$$\emptyset = FX_0 \subset FX_1 \subset FX_2 \subset FX_3 \dots$$

that control how fast the family grows. The definition is rather complicated, so we present it step-by-step, along with the notation. For any $n_i \leq n < n_{i+1}$ we write $c(n) = i$, and think of it as a counter. The sets FX_k are going to be defined such that the following hold:

1. FX_k is the union of some matched pairs of vertices.
2. For any matched edge $vw \subset FX_k$ there is an alternating path p that starts in x, lies entirely in FX_k, ends with the matched edge (in either direction) and has length at most $2k$.
3. $FX_{c(n)} \subset F_n(x)$ holds for all n when x is tough, as shown on this scheme of evolution:

$$n_0 \xrightarrow{FX_0=\emptyset} n_1 \xrightarrow{FX_1 \subset F_{n_1}(x)} n_2 \xrightarrow{FX_2 \subset F_{n_2}(x)} n_3 \xrightarrow{FX_3 \subset F_{n_3}(x)} n_4 \dots$$

Suppose we have already fixed n_k and FX_k.

Definition 6.40. Let m_k denote the smallest moment $m_k > n_k$ in which there are at most d edges leaving $FX_k \cup \{x\}$ that do not end in $B_{m_k} \setminus FX_k$. Let this set of edges be denoted by E_k. Now define FX_{k+1} to be the extension of FX_k by those matched edges in B_{m_k} that can be the last edge of an alternating path of length at most $2k$ starting from x, and lying entirely in FX_k except for its last two vertices.

$$FX_{k+1} = FX_k \cup \{v \in B_{m_k} : \exists p \in \mathfrak{A}_e(|p| \leq 2k+2; p_0 = x; \\ p_1, \dots, p_{2k} \in FX_k; p_{2k+1} = v \text{ or } p_{2k+2} = v\} \quad (6.6)$$

It is clear that this construction satisfies the first two conditions stated just above Definition 6.40, but there is no reason for FX_{k+1} to be a subset of $F_{m_k}(x)$. However, if we choose $n_{k+1} = m_k + 2k$ then the following claim implies that the third condition will be also satisfied.

Claim 6.41. *While x is tough, $FX_{k+1} \subset F_{m_k+2k}(x)$ for all k, hence $FX_{c(n)} \subset F_n(x)$ for all n.*

Proof. This is very similar to Claim 6.36. We argue by induction on k. Then we can assume that $FX_k \subset F_{m_k}$. We need to show that $FX_{k+1} \subset F_{m_k+2k+1}$. Take a matched edge $vw \subset FX_{k+1} \setminus FX_k$. By definition there is an alternating path p of length at most $2k+2$ starting in x, ending in the vw edge, and lying in FX_k. Suppose its last vertex is w. Since $vw \subset B_{m_k}$, there has to be a path q proving this, ending in the same edge, but in the opposite order: wv. Let's take the shortest such path. It has to pass through x, otherwise x would not be tough at $n = m_k + k$. Denote the part of this path between x and v by q. Now we have two paths from x. The path p ends with vw while the path q ends with wv. The length of p is at most $2k+2$, the length of q is at most $2a(x)$. We will show that some subset of q together with F_{m_k} satisfies the descendent property at $n = m_k + 2k + 1$.

Lemma 6.42. *Suppose p and q are alternating paths, both starting with a non-matched edge from the same vertex x and ending in a matched edge vw but from different directions. Then there is a subset $U \subset q$ containing v and w, such that for each vertex $z \in U$ there are two alternating paths between x and z of different length-parities, lying entirely in $U \cup p$, whose total length is at most $|q| + 2|p| - 3$.*

Before giving the proof of the lemma, let us show how this completes the proof of the claim. It is easy to see that $U \cup F_{m_k}$ satisfies the descendent property at time $m_k + 2k$. First of all, by definition, the set F_{m_k} itself satisfies it. On the other hand for any vertex in U the lemma guarantees the existence of the two alternating paths lying entirely in $U \cup p \subset U \cup F_{m_k}$, since $p \subset F_{m_k}$ by induction. The sum of the length of these two paths is at most $2a(x)+2(2k+2)-3 = 2a(x) + 4k + 1$. The age of x at $n = m_k + 2k$ is $a(x) + 2k$ and so we are done. This completes the proof of the induction step, hence the claim is true. □

Proof of Lemma 6.42. If p and q are disjoint apart from their endpoints, then the statement is obvious with $U = q$, and we even get the stronger upper bound $|q| + |p|$ on the total length of the two paths for any vertex in U. If p and q are badly intertwined, we need to be cautious. Let $x = q_0, q_1, \ldots, q_{2l} = v$ denote the vertices of q. Since both p and q are alternating paths, their intersection is necessarily a union of matched edges. For each matched edge $q_{2i-1}q_{2i}$ the path p may contain this edge, or not. The ones that are contained in p will be called double edges. For each double edge, p may contain it in the same orientation as q - these will be called good double edges, or the opposite orientation as q - these will be called bad double edges.

There are two natural partial orders on the set of matched edges of p and q. For two such edges e and f will write $e <_q f$ if e comes before f on the path q. We will write $e <_p f$ if e comes before f on p. (If one or both of the edges aren't on a given path, they are incomparable in the given order.) Now for any matched edge e on q, we define

$$Z(e) = \min_{<_p} \{f : f \geq_q e\}.$$

Note that, since the q-maximal edge vw is a double edge, $Z(e)$ is always well-defined. Also note that $Z(Z(e)) = Z(e)$. Next, let

$$f = \max_{<_q} \{e \in q : Z(e) = e \text{ is a good double edge}\},$$

and let x' be the vertex of f further away from x. If there is no such double edge, then f is not defined, and we just choose $x' = x$. Let q' be the part of q from x' to v, let p' be the part of p between x' and w, and let p'' be the part of p between x and x'. We claim that $U = q' \setminus p$ is a good candidate.

First of all, observe that $p'' \cap q' = x'$. When $x' = x$ this is obvious. Otherwise it is still true because $Z(f) = f$, which means that any edge in q' is visited by p later than f is visited by p. Second, take any matched edge $e \in q'$. By definition, $f <_q e$. Hence

$$f <_q e \leq_q Z(e) = Z(Z(e)),$$

so by construction $Z(e)$ has to be a bad double edge. Now we can exhibit the two alternating paths between x and the edge e.

From one direction we can simply reach it by going on p'' until x' and then continuing on q' until we reach e. This is a path, since $p'' \cap q' = x'$. From the other direction, start at x and go on p'' to x' and then further on p' until hitting $Z(e)$. Since $Z(e)$ is a bad double edge, we have just visited it in the 'wrong' direction on q. So we can now continue on q backwards from $Z(e)$ until we come to e. The concatenation of these two segments is still a path, since by definition of $Z(e)$, the part of p between x and $Z(e)$ is disjoint from the part of q between e and $Z(e)$. The total length of the two paths we have just exhibited is at most $2|p''| + |q'| + |p'| - 1$. The -1 comes from the fact that the vw edge is contained in both p and q, but has to be used at most once. Finally $|p'| \geq 2$ thus the total length is at most $|q| + 2|p| - 3$. □

Now let's look at the connected component of x denoted by X'. It's a (finite or countable) connected d-regular c_0-expander graph with a partial matching. Let's remove the edge containing x from the matching. Let $S' = \{x\}$ and let $X'_k = \{x\} \cup FX_k$. We have already defined the sets E_k that contain all the edges leaving X'_k not ending in B_{m_k}, hence in particular containing the once matched edge coming out of x. The sets X'_k were constructed exactly according to the rules of Definition 6.22. Clearly $|E_k| \leq d|S'|$. But since any odd set, in particular X'_k, has at least $d+1$ edges leaving it, of which at most d is forbidden, the 4th assumption of Theorem 6.24 is also satisfied. Thus it applies in this situation and implies that as long as x remains tough and $|F_n(x)| \leq |X' \setminus F_n(x)|$, we have

$$|F_n(x)| \geq |FX_{c(n)}| \geq \frac{c_0^2|S'|}{16d^4}\left(1 + \frac{c_0^3}{128d^6}\right)^{c(n)}.$$

In the countable case the $|F_n(x)| \leq |X' \setminus F_n(x)|$ condition is always satisfied and $|S'| = 1$, while in the finite case it is satisfied as long as the family doesn't occupy at least half of the graph, and $|S'| = 1/|X|$. Thus in both cases we get

Corollary 6.43. *As long as x remains tough and $||F_n(x)|| < |X|/2$,*

$$||F_n(x)|| \geq ||FX_{c(n)}|| \geq \frac{c_0^2}{16d^4}\left(1 + \frac{c_0^3}{128d^6}\right)^{c(n)},$$

where $||\cdot||$ denotes the actual size of the set in both the finite and the countable cases.

Definition 6.44. Let us say, that for such moments when $n_i \leq n < m_i$ for some i, the family of $x \in X$ is *dormant*, whereas for moments that satisfy $m_i \leq n < n_{i+1}$ the family is *active*. Let $f_n(x)$ denote the number of moments $m < n$ such that $||F_m(x)|| < c_3/\varepsilon$ and at moment m the family was active, where ε denotes the ratio of unmatched vertices in X. We will choose $c_3 = 2$ except in the case when X is finite and $\varepsilon = 2/||X||$. In this case we will choose $c_3 = 1$.

It is clear that $f_n(x) \leq c(n)^2$. Note that in the finite case either $\varepsilon \geq 4/||X||$ and thus $c_3 = 2$ and $c_3/\varepsilon \leq ||X||/2$, or $\varepsilon = 2/||X||$ and $c_3/\varepsilon = ||X||/2$. Hence families that haven't reached the size c_3/ε are not bigger than half of the graph. Hence by 6.43 we have in both the measurable and the finite case that

$$f_n(x) \leq \left(\frac{\log\left(\frac{8c_3 d^4}{\varepsilon c_0^2}\right)}{\log\left(1 + \frac{c_0^3}{128 d^6}\right)} \right)^2 \leq c_4 \left(1 + \log^2 1/\varepsilon\right) \tag{6.7}$$

for a suitably large c_4 depending only on the previous constants and d.

6.4.6 Proof of Theorem 6.21

The proof will work similarly to that of Theorem 6.24, but one has to be more careful. This time we are interested only in the measurable case, and assume that all the sets E_k of forbidden edges are empty. Thus $\tilde{H}_k$ is simply the set of vertices that can be the end-point of an odd alternating path of length at most $2k-1$ starting in S. We choose S to be half of set of unmatched vertices. Then as soon as we have $S \cap \tilde{H}_k \neq \emptyset$ or $F \cap X_n \neq \emptyset$, we have found an augmenting path.

Let

$$J(n) = |X_n| + |B_n| + \frac{1}{2}\int_X f_n(x)dx.$$

We further reintroduce the notation from the proof of Theorem 6.24. As before, we will often drop the index n, when it does not cause confusion. Let TT denote the set of tough and TM the set of not-tough vertices within $T \cup S$. The tough vertices are further classified according to their families. TB denotes the tough vertices whose families have size at least c_3/ε. For tough vertices with smaller families, TE shall denote the ones that have active families at the moment, and TG denote the ones that have dormant families at the moment. So

$$S \cup T = TM \cup TT = TM \cup (TB \cup TE \cup TG).$$

First let's take a tough vertex $x \in TG$ whose family is small and dormant. By Definition 6.40 this means, that there are at least $d+1$ edges leaving $x \cup FX_{c(n)}$ that do not end in B_n. Let $|E(x, FX_{c(n)})| = k \leq d$. Then there are $d-k$ edges leaving $x \cup FX_{c(n)}$ from x. The rest, at least $k+1$ must leave from $FX_{c(n)}$. And since these edges do not end in B_n, they actually have to leave the whole family $F_n(x)$. The only tough vertex adjacent to the family is x by Corollary 6.35, so the $k+1$ edges we have just exhibited must end in $H \cup TM \cup O$. When $k \leq d$, then $(k+1)d/(d+1) \geq k$. So we have

$$|E(F(x), TG)| = |E(F(x), x)| \leq \frac{d}{d+1}|E(F(x), H \cup TM \cup O)|.$$

Integrating over TG we get that

$$|E(B, TG)| \leq \frac{d}{d+1}|E(B, H \cup TM \cup O)|$$

For any other tough vertex we bound the number of edges between it and B by the trivial bound d. Adding this to the previous equation we get

$$|E(B, TT)| \leq d|TB| + d|TE| + \frac{d}{d+1}|E(B, H \cup TM \cup O)| \tag{6.8}$$

Now let us examine the edges running between B_n and its complement. By (6.8) we have

$$|E(B, X \setminus B) = |E(B, H \cup O \cup TM)| + |E(B, TT)| \leq \\ \leq 2|E(B, H \cup O \cup TM)| + d|TB| + d|TE|$$

and hence

$$\frac{|E(B, X \setminus B)|}{2(d+1)} \leq \frac{|E(B, H \cup TM \cup O)|}{d+1} + \frac{1}{2}(|TB| + |TE|).$$

Adding this to (6.8) then yields

$$\frac{|E(B, X \setminus B)|}{2(d+1)} + |E(B, TT)| \leq |E(B, H \cup TM \cup O)| + (d+1)(|TB| + |TE|). \tag{6.9}$$

We know that $|T| = |H|$ because of the matching, so the total degrees of $S \cup T$ is $d|S|$ more than the total degree of H. The edges between $T \cup S$ and H contribute equally to these total degrees. In the worst case there are no internal edges in H. This boils down to the following estimate.

$$|E(H, O)| + |E(H, B)| + d|S| \leq \\ \leq 2|E(T \cup S, T \cup S)| + |E(T \cup S, O)| + |E(TM, B)| + |E(TT, B)|.$$

Adding $\frac{|E(B, X \setminus B)|}{2(d+1)}$ to both sides, then using (6.9), and subtracting $|E(H, B)|$ from both sides we get

$$|E(H, O)| + \frac{|E(B, X \setminus B)|}{2(d+1)} + d|S| \leq 2|E(T \cup S, T \cup S)| + 2|E(B, TM)| + \\ + |E(B \cup T \cup S, O)| + (d+1)(|TB| + |TE|).$$

Adding $|E(B \cup T \cup S, O)|$ to both sides implies

$$|E(X_n, O)| + \frac{|E(B, X \setminus B)|}{2(d+1)} + d|S| \leq 2|E(T \cup S, T \cup S) + 2|E(B, TM)| + \\ + 2|E(B \cup T \cup S, O)| + (d+1)\left(|TB| + |TE|\right). \tag{6.10}$$

Any vertex in O_n that is adjacent to $B_n \cup T_n \cup S$ is going to be in X_{n+1}, hence

$$|E(B_n \cup T_n \cup S, O_n)| \leq d(|X_{n+1}| - |X_n|).$$

By definition, any vertex in TM_n that is adjacent to an edge coming from B_n will be part of B_{n+1} or yield an augmenting path. Also, by Lemma 6.25, any edge in $E(S \cup T_n, S \cup T_n)$ has to be adjacent to a point in $|B_{n+1}| \setminus |B_n|$ or yield an augmenting path. This implies that

$$2|E(T, T)| + 2|E(B, TM)| \leq 2d(|B_{n+1}| - |B_n|).$$

Plugging all this into (6.10) we get

$$\frac{|E(X_n, O_n)|}{d+1} + \frac{|E(B, X \setminus B)|}{2(d+1)^2} + |S| \leq \\ \leq \frac{2d}{d+1}\left(|X_{n+1}| - |X_n| + |B_{n+1}| - |B_n|\right) + |TE| + |TB| \tag{6.11}$$

By Definition 6.37, for any vertex $x \in TE_n$ we get $f_{n+1}(x) = f_n(x) + 1$, and thus

$$\int_X f_{n+1}(x)dx = \int_X f_n(x)dx + |TE|.$$

Hence the right hand side of (6.11) is at most $2(J(n+1) - J(n)) + |TB|$. Furthermore by the expander assumption we have

$$|E(X_n, O_n)| \geq c_0|X_n|(1 - |X_n|)$$

and

$$|E(B_n, X \setminus B_n)| \geq c_0|B_n|(1 - |B_n|)$$

so from (6.11) we get

$$\frac{c_0|X_n|(1-|X_n|)}{d+1} + \frac{c_0|B_n|(1-|B_n|)}{2(d+1)^2} + |S| - |TB| \leq 2(J(n+1) - J(n)) \tag{6.12}$$

Any vertex in TB has a family of size at least c_3/ε, and all these are disjoint by Claim 6.34 and contained in B. Thus we get that $|TB| \leq \varepsilon|B|/c_3 \leq \varepsilon/2 = |S|$ in the measurable case and in the finite case when $\varepsilon \geq 4/||X||$. In the finite case when $\varepsilon = 2/||X||$, then any tough vertex in TB has a family of size at least $||X||/2$, and thus there can be at most one tough vertex. We get $|TB| \leq |S|$ in all cases, and thus

$$\frac{c_0|X_n|(1-|X_n|)}{2(d+1)} + \frac{c_0|B_n|(1-|B_n|)}{4(d+1)^2} \leq J(n+1) - J(n). \tag{6.13}$$

If we could prove a similar growth estimate on the size of X_n (or B_n), then the next lemma would imply that X_n (or B_n) would grow too large in a sufficiently small number of steps, proving the existence of a short augmenting path.

Lemma 6.45. *Let $0 < a_0 < a_1 < a_2, \ldots$ be an increasing sequence of numbers. Let us fix a constant c and say that an index k is good if $a_{k+1} - a_k \geq 2ca_k(1-a_k)$ holds. Then if the number of good indices up to N is at least*

$$2\left\lceil \frac{\log(\frac{1}{2a_0})}{\log(\frac{1}{1-c})} \right\rceil,$$

then $a_N > 1 - a_0$.

Proof. Let us split the sequence into two parts. The first part will be where $a_k < 1/2$ and the second part where $a_k \geq 1/2$.

In the first part if k is a good index then $a_{k+1} \geq a_k(1+c)$. Hence $a_k \geq a_0(1+c)^{g(k)}$ where $g(k)$ denotes the number of good indices up to k. So if

$$g(k_1) \geq \left\lceil \frac{\log(\frac{1}{2a_0})}{\log(1+c)} \right\rceil$$

we must have $a_{k_1} > 1/2$, or in other words k_1 already has to be in the second part.

In the second part a good index k implies $1 - a_{k+1} \leq (1-a_k)(1-c)$, hence if N is such that

$$g(N) = g(k_1) + \left\lceil \frac{\log(\frac{1}{2a_0})}{\log(\frac{1}{1-c})} \right\rceil \leq 2\left\lceil \frac{\log(\frac{1}{2a_0})}{\log(\frac{1}{1-c})} \right\rceil$$

then we must have $1 - a_N < a_0$. □

The problem is that (6.13) doesn't directly imply such a growth estimate on either X_n or B_n because a priori the integral term in J_n could absorb any growth implied by the inequality. We need one final trick to overcome this difficulty. The idea is that we don't need X_n or B_n to grow the desired amount in one single step. If we can find a not so large K such that $|X_{n+K} - X_n| \geq 2c(|X_n|)(1-|X_n|)$, or $|B_{n+K} - B_n| \geq 2c(|B_n|)(1-|B_n|)$, we are still good. So let us fix some K, whose precise value is to be determined later, and assume that

$$|X_{n+K}| - |X_n| < \frac{c_0}{2}(|X_n|)(1-|X_n|) \text{ and } |B_{n+K}| - |B_n| < \frac{c_0}{2}(|B_n|)(1-|B_n|).$$

This means that the growth of $J(n)$ implied by (6.13) has to largely come from the $\int f_n$ term. But note that once a vertex x has a positive f-value, then it has to be tough for the rest of its life, until it becomes part of B_m for some later m, and from that point on its f-value remains constant. Hence if for some x we find that $f_{n+K}(x) > f_n(x)$, then either $x \in B_{n+K} \setminus B_n$, or $x \in TT_{n+K}$. Also by (6.7) we know that $f_{n+K}(x) - f_n(x) \leq c_4(1 + \log^2 1/\varepsilon)$. Hence we get

$$\int_X f_{n+K}(x)dx - \int_X f_n(x)dx \leq c_4(1+\log^2 1/\varepsilon)(|B_{n+K} \setminus B_n| + |TT_{n+K}|). \tag{6.14}$$

Further it is obvious that $|TT_{n+K}| < 1 - |B_{n+K}| \leq 1 - |B_n|$ and since each vertex in TT_{n+K} has a unique, non-empty family inside B_{n+K}, we also get that $|TT_{n+K}| \leq |B_{n+K}| \leq |B_n| + c_0/2|B_n|(1-|B_n|) \leq 2|B_n|$. Hence we can simply write

$$|TT_{n+K}| \leq 4|B_n|(1-|B_n|)$$

because either $|B_n|$ or $1-|B_n|$ is at least $1/2$. We also have by assumption that $|B_{n+K} \setminus B_n| \leq c_0/2|B_n|(1-|B_n|) \leq |B_n|(1-|B_n|)$. Plugging all this into (6.14) we get

$$\int_X f_{n+K}(x)dx - \int_X f_n(x)dx \leq c_4(1+\log^2 1/\varepsilon)5|B_n|(1-|B_n|), \tag{6.15}$$

and by the assumptions on the small growth of X_n and B_n we can further deduce (assuming c_4 is not really small)

$$J(n+K) - J(n) \leq (6c_4(1+\log^2 1/\varepsilon))|B_n|(1-|B_n|) + c_0/2|X_n|(1-|X_n|). \tag{6.16}$$

On the other hand we can apply (6.13) to $n, n+1, \ldots, n+K-1$. By the assumption on the small growth of X_n and B_n during this time, $|X_n|(1-|X_n|)$ and $|B_n|(1-|B_n)$ do not change too much either. More precisely we can write for any $n \leq m < n+K$ that $|X_n| \leq |X_m|$ and that $1 - |X_{n+K}| \leq 1 - |X_m|$. Also

$$(1-|X_n|) - (1-|X_{n+K}|) \leq \frac{c_0}{2}|X_n|(1-|X_n|)$$

and thus

$$1 - |X_{n+K}| \geq (1 - \frac{c_0}{2}|X_n|)(1-|X_n) \geq \frac{1-|X_n|}{2}.$$

Putting all this together we get that

$$|X_m|(1-|X_m|) \geq |X_n|(1-|X_{n+K}|) \geq \frac{1}{2}|X_n|(1-|X_n|),$$

and the exact same equation holds for B_m. Now summing (6.13) for $n, n+1, \ldots, n+K-1$ and using the last inequality, we find that

$$\frac{K}{2}\left(\frac{c_0|X_n|(1-|X_n|)}{2(d+1)} + \frac{c_0|B_n|(1-|B_n|)}{4(d+1)^2}\right) \leq J(n+K) - J(n) \tag{6.17}$$

Now choose K so large that $K > 2(d+1)$ and $c_0 K > 24c_4(d+1)^2(1+\log^2 1/\varepsilon))$, and we clearly have a contradiction between (6.16) and (6.17).

Corollary 6.46. *This implies that for any n either $|X_{n+K}| - |X_n| \geq \frac{c_0}{2}|X_n|(1 - |X_n|)$ or $|B_{n+K}| - |B_n| \geq \frac{c_0}{2}|B_n|(1 - |B_n|)$.*

Let us consider the sequences $a_n = |X_{nK+n_0}|, b_n = |B_{nK+n_0}|$. Then Corollary 6.46 implies, using the language of Lemma 6.45, that every n is a good moment for either a_n or b_n. We know that $a_0|X_{n_0}| = \varepsilon/2 > \varepsilon/6$. If we also knew that $b_0 = |B_{n_0}| \geq \varepsilon/8$, then by Lemma 6.45 we could deduce that for

$$k = n_0 + 4K \left\lceil \frac{\log(4/\varepsilon)}{\log(4/(4 - c_0))} \right\rceil$$

we have $|X_k| > 1 - \varepsilon/8 \geq 1 - \varepsilon/2$ or $|B_k| > 1 - \varepsilon/8 \geq 1 - \varepsilon/2$, either of which implies the existence of an augmenting path. All we need to do to finish the proof of Theorem 6.21 is to exhibit a not too large n_0 for which $|B_{n_0}| > \varepsilon/8$.

To this end we prove that as long as $|B_n|$ is very small, the size of X_n has to increase rapidly. Obviously $\int f_{n+1}(x) - \int f_n(x) \leq |TE| \leq |B_n|$ since every tough vertex has a nonemtpy family. Hence

$$J(n + 1) - J(n) \leq |X_{n+1}| - |X_n| + \frac{3}{2}|B_{n+1}|.$$

If $|S| \geq 4|B_n|$ then, since clearly $|TB| \leq |B_n|$, we also have $|S| - |TB| \geq 3|B_n|$ and thus by (6.12) we get

$$\frac{c_0}{2(d + 1)}|X_n|(1 - |X_n|) \leq |X_{n+1}| - |X_n|.$$

Then Lemma 6.45 implies that this cannot hold for more than

$$2 \left\lceil \frac{\log(1/\varepsilon)}{\log(\frac{1}{1-c_0/4(d+1)})} \right\rceil$$

steps. So this is a good choice for n_0. The dependence of K on $\log(1/\varepsilon)$ is quadratic, of n_0 linear, hence k is of order $O(\log^3(1/\varepsilon))$, the implied constant only depending on c_0 and d. This completes the proof of Theorem 6.21.

Chapter 7

Core percolation on complex networks

As a fundamental structural transition in complex networks, core percolation is related to a wide range of important problems. Yet, previous theoretical studies of core percolation have been focusing on the classical Erdős-Rényi random networks with Poisson degree distribution, which are quite unlike many real-world networks with scale-free or fat-tailed degree distributions. Here we show that core percolation can be analytically studied for complex networks with arbitrary degree distributions. We derive the condition for core percolation and find that purely scale-free networks have no core for any degree exponents. We show that for undirected networks if core percolation occurs then it is always continuous while for directed networks it becomes discontinuous when the in- and out-degree distributions are different. We also apply our theory to real-world directed networks and find, surprisingly, that they often have much larger core sizes as compared to random models. These findings would help us better understand the interesting interplay between the structural and dynamical properties of complex networks.

Network science has emerged as a prominent field in complex system research, which provides us a novel perspective to better understand complexity [56, 57, 58]. In the last decade considerable advances about structural and dynamical properties of complex networks have been made [59, 60, 61]. Among them, structural transitions in networks were extensively studied due to their big impacts on numerous dynamical processes on networks. Particularly interesting are the emergence of a giant connected component [62, 63, 64, 65] , k-core percolation [66, 67, 68], k-clique percolation [69, 70], and explosive percolation [71, 72, 73]. These structural transitions affect many properties of networks, e.g. robustness and resilience to breakdowns [74, 64, 75], cascading failure in interdependent networks [76, 77, 78, 79], epidemic and information spreading on socio-technical systems [80, 81, 58]. Recent work on network controllability reveals another interesting interplay between the structural and dynamical properties of complex networks [82, 83, 84]. It was found that the robustness of network controllability is closely related to the presence of the *core* in the network [82, 85]. Actually, core percolation has also been related to many other interesting problems, including conductor-insulator transitions [86, 87] and some classical combinatorial optimization problems, e.g. maximum matching [88, 89, 90] and vertex cover [91, 92, 93].

The core of a undirected network is defined as a spanned subgraph which remains in the network after the following greedy leaf removal (GLR) procedure [88, 87]: As long as the network has leaves, i.e. nodes of degree 1, choose an arbitrary leaf v_1, and its neighbor v_2, and remove them together with all the edges incident with v_2. Finally, we remove all isolated nodes. It can be proven that the resulting graph is independent of the order of removals [87]. Note that the core described above is fundamentally different from the k-core of a network. The latter is defined to be the maximal subgraph having minimum node degree of at least k,

which can be obtained by iteratively removing nodes of degree less than k. Apparently, the GLR procedure described above is more destructive than the node removal procedure used to obtain the 2-core (see Fig.7.1a). In studying the robustness of controllability for general directed networks, the GLR procedure has been extended to calculate the core of directed networks [82]. We first transform a directed network $\mathcal{G}$ to its bipartite graph representation $\mathcal{B}$ by splitting each node v into two nodes v^+ (upper) and v^- (lower), and we connect v_1^+ to v_2^- in $\mathcal{B}$ if there is a link $(v_1 \to v_2)$ in $\mathcal{G}$. The core of a directed network $\mathcal{G}$ can then be defined as the core of its corresponding bipartite graph $\mathcal{B}$ obtained by applying GLR to $\mathcal{B}$ as if $\mathcal{B}$ is a unipartite undirected network.

One can easily tell whether the core exists in two very special cases: (1) If a network has no cycles, i.e. a tree or a forest (a disjoint union of trees), then eventually all nodes will be removed, hence no core. For example, the Barabási-Albert (BA) model with parameter $m = 1$ yields a tree network, hence no core exists. (2) If a network has no leaf nodes, e.g. regular graphs with all nodes having the same degree $k > 1$ or the networks generated by the BA model with $m > 1$, then the GLR procedure will not even be initiated, hence all the nodes belong to the core.

Except those two special cases, no general rules have been proposed to predict the existence of the core for an arbitrarily complex network. Previous theoretical studies focused on undirected Erdős-Rényi (ER) random graph. It has been show that for mean degree $c \leq e = 2.7182818\ldots$, the core is small (zero asymptotically), whereas for $c > e$ the core covers a finite fraction of all the nodes [88, 87, 94]. In other words, core percolation occurs at the critical point $c* = e$. More interestingly, it has been suggested that in ER random graph core percolation coincides with the changes of the solution-space structure of the vertex cover problem [91, 93, 95], which is one of the basic NP-complete optimization problems [96]. Also, for $c \leq e$ the typical running time of an algorithm for finding the minimum vertex cover is polynomial [91, 87], while for $c > e$, one needs typically exponential running time [97]. Hence, core percolation also coincides with an "easy-hard transition" of the typical computational complexity [93, 95].

Despite the results on undirected ER random networks and the importance of understanding the intriguing interplay between core percolation and other problems, we lack a systematic study and a general theory of core percolation for both undirected and directed random networks with arbitrary degree distributions.

7.1 Analytical framework

We propose the following analytical framework to study core percolation on random networks with arbitrary degree distributions. We first categorize the nodes according to how they can be removed during the GLR procedure. We define the following categories: (1) α-removable: nodes that can become isolated (e.g. v_1 and v_2 in Fig.7.1b); (2) β-removable: nodes that can become a neighbor of a leaf (e.g. v_3 and v_5 in Fig.7.1b); (3) non-removable: nodes that cannot be removed and hence belong to the core (e.g. v_6, v_7 and v_8 in Fig.7.1b). While the core is independent of the order the leaves are removed [87], the specific way a node is removed may depend on this order, but it can be proven that no node can be both α-removable and β-removable at the same time. Now we consider an uncorrelated random network with arbitrary degree distribution $P(k)$ [65, 98]. Assuming that in each removable category the removal of a random node can be made locally, we can determine the category of a node v in a network $\mathcal{G}$ by the categories of its neighbors in $\mathcal{G} \setminus v$, i.e. the subgraph of $\mathcal{G}$ with node v and all its edges removed, using the following rules: (1) α-removable: all neighbors are β-removable; (2) β-removable: at least

one neighbor is α-removable; (3) non-removable: no neighbor is α-removable, and at least two neighbors are not β-removable.

Let α and β denote the probability that a random neighbor of a random node v in a network $\mathcal{G}$ is α-removable and β-removable in $\mathcal{G} \setminus v$, respectively. We can derive two self-consistent equations about α and β

$$\alpha = \sum_{k=1}^{\infty} Q(k)\beta^{k-1} = A(1-\beta), \tag{7.1}$$

$$1-\beta = \sum_{k=1}^{\infty} Q(k)(1-\alpha)^{k-1} = A(\alpha) \tag{7.2}$$

where $Q(k) \equiv kP(k)/c$ is the degree distribution for the node at a random end of a randomly chosen edge, $c \equiv \sum_{k=0}^{\infty} kP(k)$ is the mean degree, and $A(x) \equiv \sum_{k=0}^{\infty} Q(k+1)(1-x)^k$. These two equations indicate that α satisfies $x = A\big(A(x)\big)$. It can be shown that α is the smallest fixpoint of $A\big(A(x)\big)$, i.e. the smallest root of the function $f(x) \equiv A\big(A(x)\big) - x$.

The expected fraction of non-removable nodes, i.e. the normalized core size ($n_{\text{core}} \equiv N_{\text{core}}/N$), can then be calculated:

$$n_{\text{core}} = \sum_{k=0}^{\infty} P(k) \sum_{s=2}^{k} \binom{k}{s} \beta^{k-s}(1-\beta-\alpha)^s, \tag{7.3}$$

which can be simplied in terms of $G(x) \equiv \sum_{k=0}^{\infty} P(k)x^k$, i.e. the generating function of the degree distribution $P(k)$. The final result is given by

$$n_{\text{core}} = G(1-\alpha) - G(\beta) - c\,(1-\beta-\alpha)\,\alpha. \tag{7.4}$$

For Erdős-Rényi random networks, $G(x) = e^{-c(1-x)} = A(1-x)$, Eq.7.4 can be further simplified as $n_{\text{core}} = (1-\beta-\alpha)(1-c\alpha)$, confirming previous results [87, 94].

The normalized number of edges in the core ($l_{\text{core}} \equiv L_{\text{core}}/N$) can also be calculated in terms of α and β. Consider a uniform random edge, which remains in the core if and only if both of its endpoints are non-removable without removing the edge. The probability of one endpoint being non-removable without removing the edge is $1-\alpha-\beta$, and for the two endpoints the probabilities are independent. Therefore, the expected normalized number of edges in the core is

$$l_{\text{core}} = \frac{c}{2}\,(1-\alpha-\beta)^2. \tag{7.5}$$

with $c/2 = L/N$ the normalized number of edges in the network. Clearly, both $n_{\text{core}} > 0$ and $l_{\text{core}} > 0$ if and only if $1-\beta-\alpha > 0$.

Now we consider directed networks $\mathcal{G}$ with given in- and out-degree distributions, denoted by $P^-(k)$ and $P^+(k)$, respectively. Let c denote the mean degree of each partition in the bipartite graph representation $\mathcal{B}$ of the directed network $\mathcal{G}$, i.e. the mean in-degree (or out-degree) of $\mathcal{G}$. Define $Q^{\pm}(k) \equiv kP^{\pm}(k)/c$, which is the degree distribution of the upper or lower end, respectively, of a random edge in $\mathcal{B}$. Define $A^{\pm}(x) \equiv \sum_{k=0}^{\infty} Q^{\pm}(k+1)(1-x)^k$. Then the same argument as we used in the undirected case gives that

$$\alpha^{\pm} = A^{\pm}(1-\beta^{\mp}), \tag{7.6}$$

$$1-\beta^{\pm} = A^{\pm}(\alpha^{\mp}) \tag{7.7}$$

and $\alpha^{\pm}$ is the smallest fixpoint of $A^{\pm}\big(A^{\mp}(x)\big)$. Now we can calculate the size of the core for each partition in $\mathcal{B}$ as

$$n_{\rm core}^{\pm} = \sum_{k=0}^{\infty} P^{\pm}(k) \sum_{s=2}^{k} \binom{k}{s} (\beta^{\mp})^{k-s} (1-\beta^{\mp}-\alpha^{\mp})^{s} \tag{7.8}$$

and we define the size of the core in the directed network $\mathcal{G}$ as

$$n_{\rm core} = (n_{\rm core}^{+} + n_{\rm core}^{-})/2. \tag{7.9}$$

The normalized number of edges in the core can also be calculated

$$l_{\rm core} = c\,(1-\alpha^{+}-\beta^{+})(1-\alpha^{-}-\beta^{-}). \tag{7.10}$$

7.2 Condition for core percolation

It is easy to see that the core in a undirected network with degree distribution $P(k)$ is the very same as in a directed network with the same out- and in-degree distributions, i.e. $P^{+}(k) = P^{-}(k) = P(k)$. Therefore we can deal with directed network for generality. As $n_{\rm core}$ is a continuous function of $\alpha^{\pm}$, we focus on $\alpha^{\pm}$, which is the smallest root of the function $f^{\pm}(x) \equiv A^{\pm}\big(A^{\mp}(x)\big) - x$. There are several interesting facts about the function $f^{\pm}(x)$. First of all, since $A^{\pm}(x)$ is a monotonically decreasing function for $x \in [0,1]$ and $A^{\pm}(0) = 1$ is the maximum (see Figs.7.2, 7.3), we have $f^{\pm}(0) > 0$ and $f^{\pm}(1) < 0$ (see Fig.7.3c,d). Consequently, the number of roots (with multiplicity) of $f^{\pm}(x)$ in $[0,1]$ is odd, and numerical calculations suggest that this number is either 1 or 3 (see Figs.7.2, 7.3). Secondly, if $f^{\pm}(x_0) = 0$ then $f^{\mp}\big(A^{\mp}(x_0)\big) = 0$, which means $A^{\mp}(x)$ transforms the roots of $f^{\pm}(x)$ to the roots of $f^{\mp}(x)$. This also suggests that $f^{\pm}(x)$ *always* has a trivial root $\alpha^{\pm} = A^{\pm}(\alpha^{\mp}) = 1-\beta^{\pm}$. (For undirected networks, $f(x)$ *always* has a trivial root $\alpha = A(\alpha) = 1-\beta$.) Since $A^{\mp}(x)$ is a monotonically decreasing function and $\alpha^{\pm}$ is the smallest root of $f^{\pm}(x)$, $A^{\mp}(\alpha^{\pm}) = 1-\beta^{\mp}$ is therefore the largest root of $f^{\mp}(x)$. Hence $1-\beta^{\pm}-\alpha^{\pm}$ is the difference between the largest and the smallest roots of $f^{\pm}(x)$ (see Fig.7.2). Consequently, if $f^{\pm}(x)$ has only one root (which then must be the trivial root $\alpha^{\pm} = A^{\pm}(\alpha^{\mp}) = 1-\beta^{\pm}$), then $1-\beta^{\pm}-\alpha^{\pm} = 0$. According to Eq.7.8, this implies that there is no core. On the other hand, if multiple roots exist and they are different then $1-\beta^{\pm}-\alpha^{\pm} > 0$, and the core will develop.

We apply the above condition to the following random undirected networks with specific degree distributions [65]. (1) Erdős-Rényi (ER) [62, 63] networks with Poisson degree distribution $P(k) = e^{-c}c^{k}/k!$, $A(x) = e^{-cx}$ and $f(x) = \exp(-ce^{-cx}) - x$. As shown in Fig.7.3a, the core percolation occurs at $c = c^{*} = e$, which agrees with previous theoretical results [88, 87, 94]. (2) Exponentially distributed graphs with $P(k) = (1-e^{-1/\kappa})e^{-k/\kappa}$ and mean degree $c = e^{-1/\kappa}/(1-e^{-1/\kappa})$. We find that core percolation occurs at $c = c^{*} = 4$. (3) Purely power-law distributed networks with $P(k) = k^{-\gamma}/\zeta(\gamma)$ for $k \geq 1$, $\gamma > 2$ and $\zeta(\gamma)$ the Riemann ζ function. We find that $f(x)$ has no multiple roots and hence $n_{\rm core} = 0$ for all $\gamma > 2$. In other words, for purely scale-free (SF) networks, the core does not exist. (4) Power-law distributed networks with exponential degree cutoff, i.e. $P(k) = \frac{k^{-\gamma}\,e^{-k/\kappa}}{{\rm Li}_{\gamma}(e^{-1/\kappa})}$ for $k \geq 1$ with ${\rm Li}_{n}(x)$ the nth polylogarithm of x. We find that $n_{\rm core} = 0$ for $\gamma > \gamma_{\rm c}(\kappa)$, and the threshold value $\gamma_{\rm c}(\kappa)$ approaches 1 as κ increases. Hence, for SF networks with exponential degree cutoff the core still does not exist for all $\gamma > 1$. (5) Asymptotically SF networks generated by the static model with $P(k) = \frac{[\frac{\xi}{2}(1-\xi)]^{1/\xi}}{\xi} \frac{\Gamma(k-1/\xi, \frac{\xi}{2}[1-\xi])}{\Gamma(k+1)}$, where $\Gamma(s)$ is the gamma function and $\Gamma(s,x)$ the upper

incomplete gamma function [99, 100, 101]. In the large k limit, $P(k) \sim k^{-(1+\frac{1}{\xi})} = k^{-\gamma}$ where $\gamma = 1 + \frac{1}{\xi} > 2$. For small k, $P(k)$ deviates significantly from the power-law distribution [100] and there are much fewer small-degree nodes than the purely scale-free networks, which results in a drastically different core percolation behavior.

Hereafter, we systematically study the net effect of adding more links (i.e. increasing mean degree c, yet without changing other parameters in $P(k)$) on core percolation. ER networks and the asymptotically SF networks generated by the static model naturally serve this purpose, since their mean-degree is an independent and explicit tuning parameter.

7.3 Nature of core percolation

We observed that if the mean degree c is small, then $f^{\pm}(x)$ has one root, but if c is large, $f^{\pm}(x)$ has three roots (see Figs.7.2, 7.3). At the critical point $c = c^*$, the number of roots jumps from 1 to 3 by the appearance of one new root with multiplicity 2. (Note that $f^{\pm}(x)$ cannot immediately intersect the x-axis at two new points, but it touches first.) This explains why the core percolation occurs at $c = c^*$.

According to the transformation from the roots of $f^{\pm}(x)$ to the roots of $f^{\mp}(x)$ through $A^{\mp}(x)$, for either $f^{+}(x)$ or $f^{-}(x)$ (depending on the details of $P^{+}(k)$ and $P^{-}(k)$) its new root at $c = c^*$ is smaller than its original root; and for either $f^{-}(x)$ or $f^{+}(x)$ the new root at $c = c^*$ is larger than the original root; or there is a degenerate case when this new root is the same as the original root for both $f^{+}(x)$ and $f^{-}(x)$. For example, for directed asymptotically SF networks generated by the static model with $\gamma_{\rm in} = 2.7, \gamma_{\rm out} = 3.0$, the new root (marked as green dot) of $f^{+}(x)$ at $c = c^*$ is smaller than the original root (green square) of $f^{+}(x)$ (see Fig.7.3c), and the new root (green square) of $f^{-}(x)$ at $c = c^*$ is larger than the original root (green circle) of $f^{-}(x)$ (see Fig.7.3d). In other words, at the critical point, for either $f^{+}(x)$ or $f^{-}(x)$, its smallest two roots are the same, and for the other function (either $f^{-}(x)$ or $f^{+}(x)$), its largest two roots are the same (see Fig.7.3c,d). While for directed networks with $P^{+}(k) = P^{-}(k) = P(k)$, i.e. the degenerate case, we have $f^{+}(x) = f^{-}(x) = f(x)$, and the new root of $f(x)$ at $c = c^*$ has to be the same as the original root of $f(x)$, i.e. all three roots must be the same (see Fig.7.3a). Therefore at the critical point, unless in the degenerate case, α^{+} together with β^{-} (or α^{-} together with β^{+}) decrease discontinuously, which implies a discontinuous transition in the core size. To sum up, in the degenerate case that $P^{+}(k) = P^{-}(k) = P(k)$ core percolation is continuous, but for general non-degenerate case $P^{+}(k) \neq P^{-}(k)$, we have a discontinuous transition in both $n_{\rm core}$ and $l_{\rm core}$. These results are clearly shown in Fig.7.3b,e.

At the critical point c^*, $f^{\pm}(x)$ touches the x-axis at its new root (see Fig.7.3c,d), hence we have either $f^{+}(\alpha^{+}) = (f^{+})'(\alpha^{+}) = 0$ (or $f^{-}(1-\beta^{-}) = (f^{-})'(1-\beta^{-}) = 0$), which enable us to calculate the core percolation threshold c^*. In the degenerate case, if $c \leq c^*$ then $f(\alpha) = f'(\alpha) = 0$ can be further simplified as $A(\alpha) = \alpha$ and $[A'(\alpha)]^2 = 1$. The results of c^* for ER and SF networks generated by the static model are shown in Fig.7.4a.

The discontinuity in $n_{\rm core}$ and $l_{\rm core}$ at c^*, denoted by $\Delta_{\rm n}$ and $\Delta_{\rm l}$ respectively, can also be calculated

$$\Delta_{\rm n} = \frac{1}{2}\left(\Delta_{\rm n}^{+} + \Delta_{\rm n}^{-}\right) \tag{7.11}$$

$$\Delta_{\rm l} = c^*(1-\beta^{-,*}-\alpha^{-,*})(1-\beta^{+,*}-\alpha^{+,*}) \tag{7.12}$$

with $\Delta_{\rm n}^{\pm} \equiv G^{\pm}(1-\alpha^{\mp,*}) - G^{\pm}(\beta^{\mp,*}) - c^*\left(1-\beta^{\mp,*}-\alpha^{\mp,*}\right)\alpha^{\pm,*}$. The results of $\Delta_{\rm n}$ for ER and SF networks generated by the static model are shown in Fig.7.4b. We find that $\Delta_{\rm n} \to 0$ as $\gamma_{\rm in} \to \gamma_{\rm out}$, consistent with the result obtained above that core percolation is continuous for

undirected networks or directed networks with $P^+(k) = P^-(k)$. We also find that Δ_n increases as the differences between γ_{in} and γ_{out} increases.

We can further show that in the general non-degenerate case, core percolation is actually a hybrid phase transition [102, 67, 68], i.e. n_{core} (or l_{core}) has a jump at the critical point as at a first-order phase transition but also has a critical singularity as at a continuous transition. The results are summarized here: in the critical regime $\epsilon = c - c^* \to 0^+$

$$n_{core} - \Delta_n \sim (c - c^*)^\eta \tag{7.13}$$
$$l_{core} - \Delta_l \sim (c - c^*)^\theta \tag{7.14}$$

with the critical exponents $\eta = \theta = \frac{1}{2}$. Our calculations do not use any specific functional form of $A^\pm(x)$. Instead, we only assume that they are continuous functions of the mean degree c. Interestingly, in the degenerate or undirected case, one has a continuous phase transition ($\Delta_n = \Delta_l = 0$) but with a completely different set of critical exponents: $\eta' = \theta' = 1$ [87].

7.4 Numerical results

We check our analytical results with extensive numerical calculations by performing the GLR procedure on finite discrete networks generated by the static model [99, 100, 101]. Fig.7.5a and 7.5b show n_{core} and l_{core} (in symbols) for undirected ER networks and asymptotically SF networks with different degree exponents. For comparison, analytical results for infinite large networks are also shown (in lines). Clearly, core percolation is continuous in this case. This is fundamentally different from the $k \geq 3$-core percolation, which becomes discontinuous for ER networks and SF networks with $\gamma > 3$ [66, 67].

Fig.7.5c and 7.5d show the results of n_{core} and l_{core} for directed networks. For directed networks with the same in- and out-degree distributions, e.g. directed ER networks or directed SF networks with $\gamma_{in} = \gamma_{out}$ generated by the static model, the core percolation is still continuous. But for directed networks with different in- and out-degree distributions, e.g. directed SF networks with $\gamma_{in} \neq \gamma_{out}$ generated by the static model, the core percolation looks discontinuous. The discontinuity in n_{core} (or l_{core}) increases as the difference between γ_{in} and γ_{out} increases (see Fig.7.5e,f).

7.5 Real networks

We also apply our theory to real-world networks with known degree distributions. In Fig.7.6 we demonstrate that in some cases our analytical results calculated from Eqs.7.4, 7.5 (or Eqs.7.9, 7.10) with degree distribution as the only input predict with surprising accuracy the core size of real networks. Yet, in other cases there is a noticeable difference between theory and reality, which suggests the presence of extra structure in the real-world networks that is not captured by the degree distribution. In particular we find that almost all the directed real-world networks have larger core sizes than the theoretical predictions (see Fig.7.6a,b). In other words, those networks are "overcored". While if we treat those networks as undirected ones, their core sizes deviate from our theory in a more complicated manner. The effects of higher order correlations (e.g. degree correlations [103], clustering [104], loop structure [105] and modularity [106]) may play very important roles to explain the discrepancy between theory and reality.

7.6 Conclusion

In sum, we analytically solve the core percolation problem in both undirected and directed random networks with arbitrary degree distributions. We show the condition for core percolation. We find it is continuous in undirected networks (if it occurs), while it becomes discontinuous or hybrid in directed networks unless the in- and out-degree distributions are the same. Within each case, the critical exponents associated with the critical singularity are universal for random networks with arbitrary degree distributions parameterized continuously in mean degree. But the two cases have totally different sets of critical exponents. These results vividly illustrate that core percolation is a fundamental structural transition in complex networks and its implication on other problems, e.g. conductor-insulator transitions, combinatorial optimization problems, and network controllability issue, deserves further exploration. The analytical framework presented here also raises a number of questions, answers to which would further improve our understanding of core percolation on complex real-world networks. For example, we focused on uncorrelated random networks and leave the systematic studies of the effects of higher order correlations as future work.

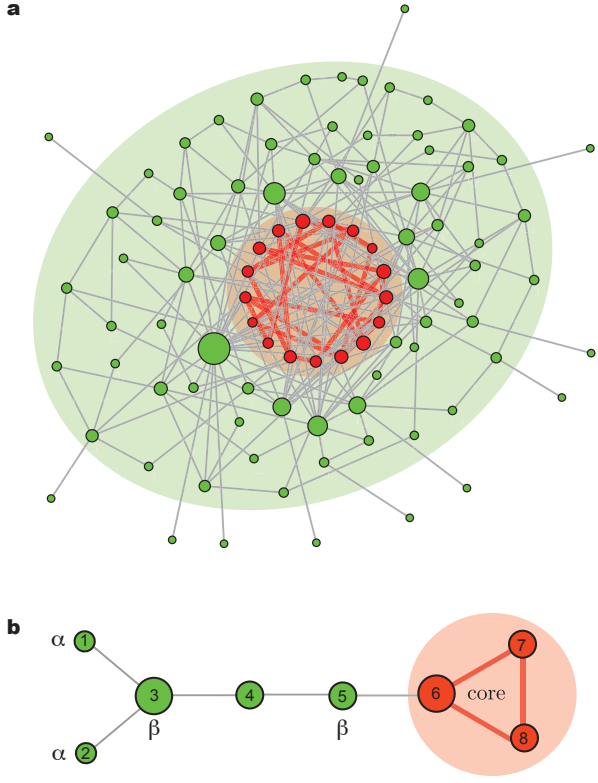


Figure 7.1: **The core of a small network. a**, The core (highlighted in red) obtained after the greedy leaf removal procedure is fundamentally different from the 2-core (highlighted in green) obtained by iteratively removing nodes of degree less than 2. The 2-core contains the core, whereas the opposite is not true. Size of nodes are roughly proportional to the degree of nodes. **b**, Removal categories of nodes according to how they can be removed during the greedy leaf removal procedure. Red nodes are non-removable, i.e. they belong to the core. Green nodes are removable: nodes v_1 and v_2 are α-removable; nodes v_3 and v_5 are β-removable. White node v_4 is removable but it is neither α-removable nor β-removable. Node v_5 is β-removable because v_4 will become a leaf node after removing node v_1 (or v_2) together with v_3.

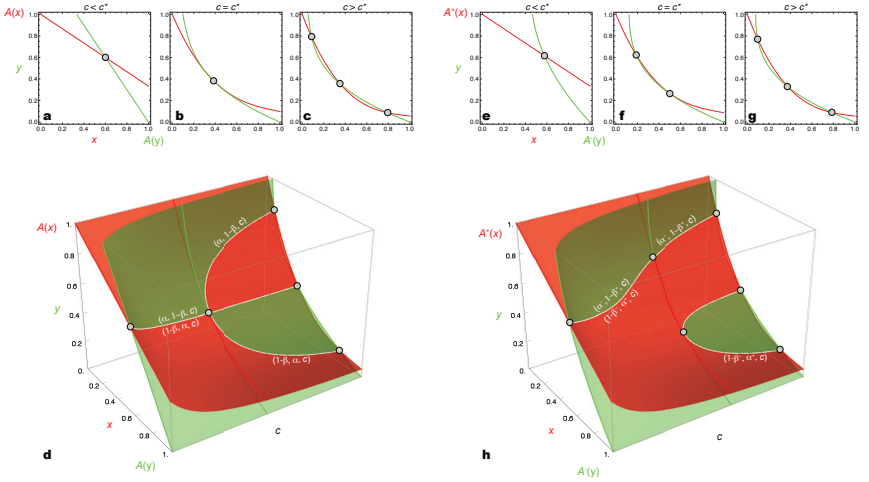


Figure 7.2: **Graphical solution of the self-consistent equations.** **a-d**, For undirected networks, the function $A(x)$ transforms the roots of $f(x)$ to the roots of the same function $f(x)$. The graphical solution of $f(x) = A\big(A(x)\big) - x = 0$ is best illustrated by plotting the two curves $A(x)$ vs. x (in red) and y vs. $A(y)$ (in green) in the same coordinate system. The coordinates of the intersection point(s) of the two curves give the solution(s) of $f(x) = 0$. In **a**, **b**, and **c**, we show the graphical solutions for $c <, =$, and $> c^*$, respectively. **d**, By drawing the two curves ($A(x)$ vs. x) and (y vs. $A(y)$) at different mean degrees c, we get two surfaces. The intersection curve of the two surfaces yields the solutions of $f(x) = 0$ at different c values. For $c < c^*$, the intersection curve has one branch given by $(\alpha, 1-\beta, c) = (1-\beta, \alpha, c)$. For $c > c^*$, the intersection curve has three branches. The top and bottom branches are given by $(\alpha, 1-\beta, c)$ and $(1-\beta, \alpha, c)$, respectively. **e-h**, For directed networks, $A^{\pm}(x)$ transforms the roots of $f^{\mp}(x)$ to the roots of $f^{\pm}(x)$. The graphical solution of $f^{\pm}(x) = A^{\pm}(A^{\mp}(x)) - x = 0$ can be illustrated by plotting $A^{+}(x)$ vs. x (in red) and y vs. $A^{-}(y)$ (in green) in the same coordinate system. The x-coordinate (or y-coordinate) of the intersection point(s) of the two curves give the solution(s) of the equation $f^{-}(x) = 0$ (or $f^{+}(x) = 0$, respectively). In **e**, **f**, and **g**, we show the graphical solutions for $c <, =$, and $> c^*$, respectively. **h**, By drawing the two curves ($A^{+}(x)$ vs. x) and (y vs. $A^{-}(y)$) at different mean degrees c, we get two surfaces. The intersection curve of the two surfaces yields the solutions of $f^{\pm}(x) = 0$ at different c values. For $c < c^*$, the intersection curve has one branch given by $(\alpha^{-}, 1-\beta^{+}, c) = (1-\beta^{-}, \alpha^{+}, c)$. For $c > c^*$, the intersection curve has three branches. The top and bottom branches are given by $(\alpha^{-}, 1-\beta^{+}, c)$ and $(1-\beta^{-}, \alpha^{+}, c)$, respectively.

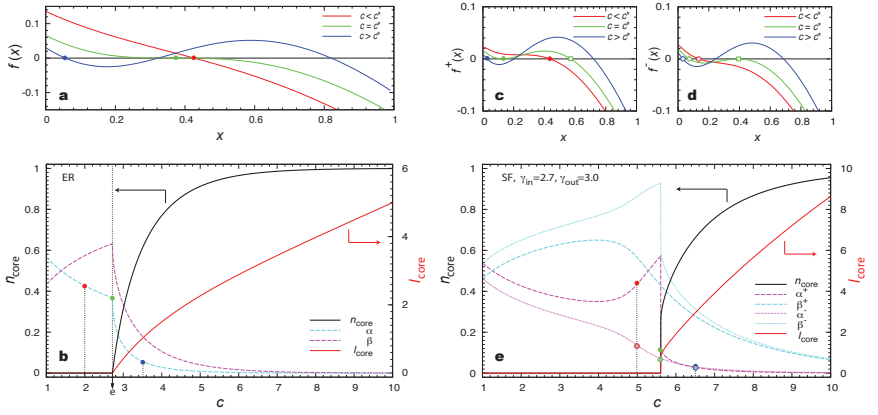


Figure 7.3: **Analytical solution of the core percolation. a-b**, Undirected Erdős-Rényi (ER) random networks. **a**, α is the smallest root of the function $f(x) \equiv A\big(A(x)\big) - x$, represented by red, green, and blue dots for $c <, =,$ and $> c^* = e$, respectively. **b**, $\alpha, \beta, n_{\rm core}$ and $l_{\rm core}$ as functions of the mean degree c. **c-e**, Directed asymptotically scale-free (SF) random networks generated by the static model. Both the in-degree and out-degree distributions of the networks are scale-free with degree exponents $\gamma_{\rm in} = 2.7$ and $\gamma_{\rm out} = 3.0$. **c, d**, $\alpha^{\pm}$ is the smallest root of the function $f^{\pm}(x) \equiv A^{\pm}\big(A^{\mp}(x)\big) - x$, represented by red, green, and blue dots for $c <, =,$ and $> c^* \simeq 11.2$, respectively. **e**, $\alpha^{\pm}, \beta^{\pm}, n_{\rm core}$ and $l_{\rm core}$ as functions of the mean degree c. The jumps in α^+ and β^- result in the jumps in $n_{\rm core}$ and $l_{\rm core}$, hence the first-order core percolation occurs.

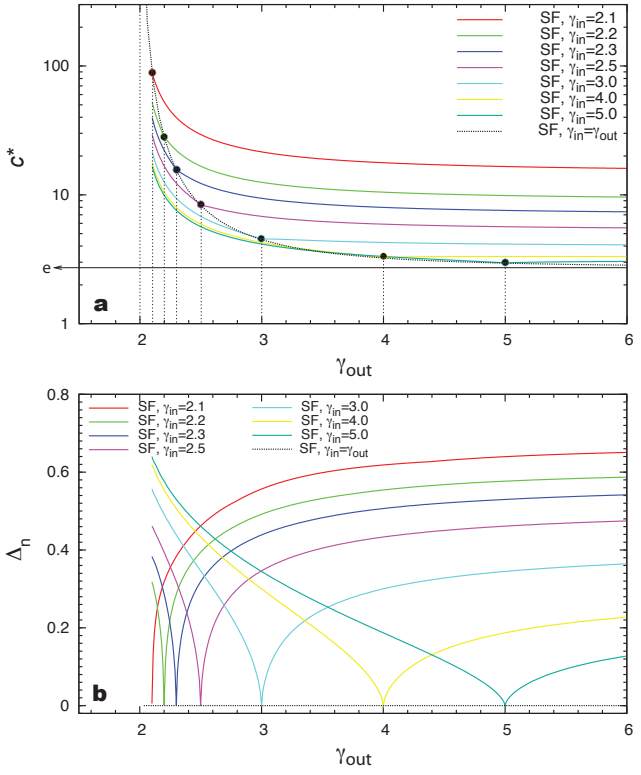


Figure 7.4: **Threshold and discontinuity of core percolation.** **a**, Analytical solution of the core percolation threshold c^* calculated by solving $f^{\pm}(x) = f^{\pm\prime}(x) = 0$ for model networks. For ER networks, $c^* = e$. For undirected asymptotically SF networks generated by the static model, $c^* \to \infty$ as $\gamma \to 2$, and and $c^* \to e$ as $\gamma \to \infty$. **b**, The discontinuity $\Delta_{\rm n}$ in $n_{\rm core}$ at $c = c^*$ for model networks. For undirected or directed networks with $P^+(k) = P^-(k)$, $\Delta_{\rm n} = 0$. For directed network, $\Delta_{\rm n}$ increases as the difference between the in- and out-degree distributions (quantified by the difference between the degree exponents $\gamma_{\rm in}$ and $\gamma_{\rm out}$) increases.

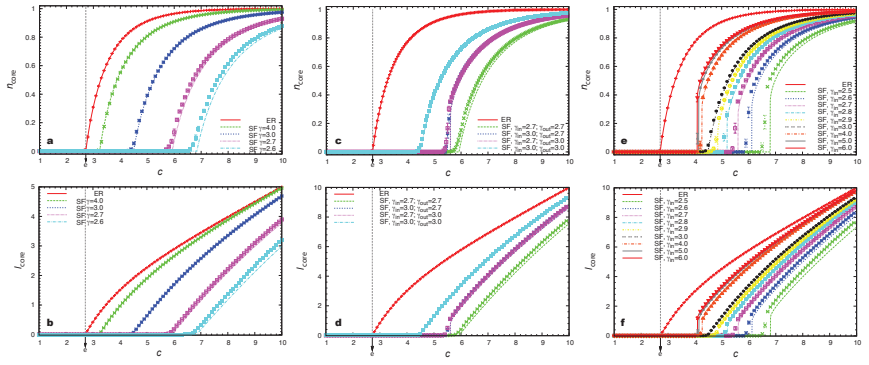


Figure 7.5: **Core percolation in random networks.** Symbols are numerical results calculated from the GLR procedure on finite discrete networks constructed with the static model [99] with $N = 10^5$. The numerical results are averaged over 20 realizations with error bars defined as s.e.m. Lines are analytical results for infinite large system ($N \to \infty$) calculated from Eq.7.4 and 7.5 for undirected networks or Eq.7.9 and 7.10 for directed networks. Finite size effects are more discernable for $\gamma \to 2$, which can be eliminated by imposing degree cutoff in constructing the SF networks [107, 108]. **a-b**, The normalized core size ($n_{\rm core} = N_{\rm core}/N$) and the normalized number of edges in the core ($l_{\rm core} = L_{\rm core}/N$) for undirected model networks: Erdős-Rényi (ER) and asymptotically scale-free (SF) with different values of γ. For both model networks, the core percolation is continuous, which is fundamentally different from the $k \geq 3$-core percolation, which becomes discontinuous for ER networks and SF networks with $\gamma > 3$ [66, 67]. **c-d**, $n_{\rm core}$ and $l_{\rm core}$ for directed ER and asymptotically SF model networks. The core percolation is continuous if the out- and in-degree distributions are the same ($P^+(k) = P^=(k)$) while it becomes discontinuous if $P^+(k) \neq P^=(k)$. **c-d**, For directed SF networks with fixed $\gamma_{\rm out} = 3.0$, by tuning $\gamma_{\rm in}$ we see that the discontinuity in both $n_{\rm core}$ and $l_{\rm core}$ become larger as the difference between $\gamma_{\rm in}$ and $\gamma_{\rm out}$ increases.

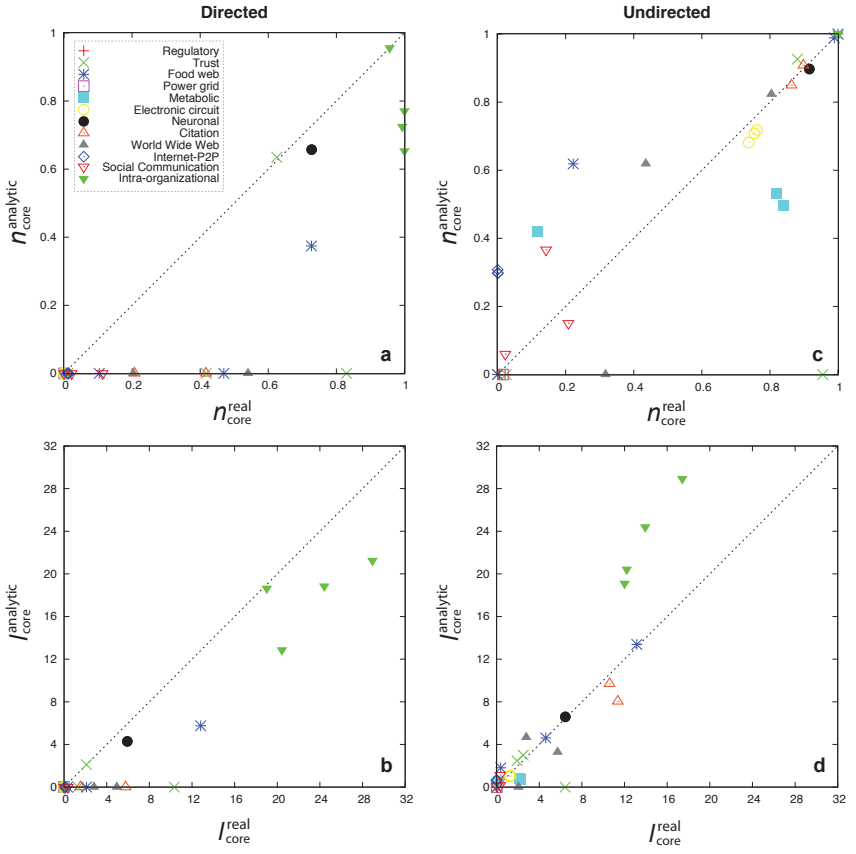


Figure 7.6: **Normalized core size for real networks, compared with analytical predictions.** All the real networks considered here are directed. For data sources and references, see Ref. [82] Supplementary Information Sec.VI. **a-b**, By applying the GLR procedure we yield $n_{\text{core}}^{\text{real}}$ and $l_{\text{core}}^{\text{real}}$. Using Eq.7.9 and Eq.7.10 with out- and in-degree distributions ($P^{+}(k)$ and $P^{-}(k)$) as the only inputs, we obtain $n_{\text{core}}^{\text{analytic}}$ and $l_{\text{core}}^{\text{analytic}}$. **c-d** By ignoring the direction of the edges, we can treat the original directed networks as undirected ones and apply the GLR procedure to get $n_{\text{core}}^{\text{real}}$ and $l_{\text{core}}^{\text{real}}$. Similarly, we can calculate $n_{\text{core}}^{\text{analytic}}$ and $l_{\text{core}}^{\text{analytic}}$ by using Eq.7.4 and Eq.7.5 with the degree distribution $P(k)$ as the only input.

Chapter 8

Positive graphs

We study "positive" graphs that have a nonnegative homomorphism number into every edge-weighted graph (where the edgeweights may be negative). We conjecture that all positive graphs can be obtained by taking two copies of an arbitrary simple graph and gluing them together along an independent set of nodes. We prove the conjecture for various classes of graphs including all trees. We prove a number of properties of positive graphs, including the fact that they have a homomorphic image which has at least half the original number of nodes but in which every edge has an even number of pre-images. The results, combined with a computer program, imply that the conjecture is true for all graphs up to 9 nodes.

8.1 Problem description

Let G and H be two simple graphs. A *homomorphism* $G \to H$ is a map $V(G) \to V(H)$ that preserves adjacency. We denote by $\hom(G, H)$ the number of homomorphisms $G \to H$. We extend this definition to graphs H whose edges are weighted by real numbers β_{ij} $(i, j \in V(H))$:

$$\hom(G, H) = \sum_{\varphi:\ V(G)\to V(H)} \prod_{ij\in E(H)} \beta_{\phi(i)\phi(j)}.$$

(One could extend it further by allowing nodeweights, and also by allowing weights in G. Positive nodeweights in H would not give anything new; whether we get anything interesting through weighting G is not investigated in this paper.)

We call the graph G *positive* if $\hom(G, H) \geq 0$ for every edge-weighted graph H (where the edgeweights may be negative). It would be interesting to characterize these graphs; in this paper we offer a conjecture and line up supporting evidence.

We call a graph *symmetric*, if its vertices can be partitioned into three sets (S, A, B) so that S is an independent set, there is no edge between A and B, and there exists an isomorphism between the subgraphs spanned by $S \cup A$ and $S \cup B$ which fixes S.

Conjecture 8.1. *A graph G is positive if and only if it is symmetric.*

There is an analytic definition for graph positivity, which is sometimes more convenient to work with. A *kernel* is a symmetric bounded measurable function $[0, 1]^2 \to \mathbb{R}$. The *weight* of a map $p \in [0, 1]^{V(G)}$ is defined as

$$\hom(G, W, p) = \prod_{e\in E} W\big(p(e)\big) = \prod_{(a,b)\in E} W\big(p(a), p(b)\big).$$

The *homomorphism density* of a graph $G = (V, E)$ in a kernel W is defined as the expected weight of a random map:

$$t(G, W) = \int_{[0,1]^V} \hom(G, W, p)dp = \int_{[0,1]^V} \prod_{e\in E} W\big(p(e)\big)dp. \tag{8.1}$$

Graphs with real edge weights can be considered as kernels in a natural way: Let H be a looped-simple graph with edge weights β_{ij}; assume that $V(H) = [n] = \{1, \dots, n\}$. Split the interval $[0, 1]$ into n intervals $J_1, \dots, J_n$ of equal length, and define

$$W_H(x, y) = \beta_{ij} \quad \text{for} \quad x \in J_i,\ y \in J_j.$$

Then it is easy to check that for every simple graph G and edge-weighted graph H, we have $t(G, W_H) = t(G, H)$, where $t(G, H)$ is a normalized version of homomorphism numbers between finite graphs:

$$t(G, H) = \frac{\hom(G, H)}{|V(H)|^{|V(G)|}}.$$

(For two simple graph G and H, $t(G, H)$ is the probability that a random map $V(G) \to V(H)$ is a homomorphism.)

It follows from the theory of graph limits [10, 44] that positive graphs can be equivalently be defined by the property that $t(G, W) \geq 0$ for every kernel W. We can also go in the other direction: a simple graph G is positive if and only if $t(G, H) \geq 0$ for every edge-weighted graph with edgeweights ± 1.

Hatami [30] studied "norming" graphs G, for which the functional $W \mapsto t(G, W)^{|E(G)|}$ is a norm on the space of kernels. Positivity is clearly a necessary condition for this (it is far from being sufficient, however). We don't know whether our Conjecture can be proved for norming graphs.

8.2 Results

In this section, we state our results (and prove those with simpler proofs). First, let us note that the "if" part of the conjecture is easy.

Lemma 8.2. *If a graph G is symmetric, then it is positive.*

Proof.

$$t(G, W) \overset{(8.1)}{=} \int_{[0,1]^V} \prod_{e\in E} W\big(p(e)\big)\, dp = \int_{[0,1]^V} \Big(\prod_{e\in S\cup A} W\big(p(e)\big)\Big) \cdot \Big(\prod_{e\in S\cup B} W\big(p(e)\big)\Big)\, dp$$

$$= \int_{[0,1]^S} \Big(\int_{[0,1]^A} \prod_{e\in S\cup A} W\big(p(e)\big)dp_A\Big) \cdot \Big(\int_{[0,1]^B} \prod_{e\in S\cup B} W\big(p(e)\big)dp_B\Big)dp_S$$

$$= \int_{[0,1]^S} \Big(\int_{[0,1]^A} \prod_{e\in S\cup A} W\big(p(e)\big)dp_A\Big)^2 dp_S \geq \int_{[0,1]^S} 0 = 0. \qquad \square$$

In the reverse direction, we only have partial results. We are going to prove that the conjecture is true for trees (Corollary 8.17), and for all graphs up to 9 nodes (see Section 8.5).

We state and prove a number of properties if positive graphs. Each of these is of course satisfied by symmetric graphs.

Lemma 8.3. *If G is positive, then G has an even number of edges.*

Proof. Otherwise $t(G,-1) = -1$. □

We call a homomorphism *even* if the preimage of each edge is has even cardinality.

Lemma 8.4. *If G is positive, then there exists an even homomorphism of G into itself.*

Proof. Let H be obtained from G by a random weighting of its edges, and let ϕ be a random map $V(G) \to V(H)$. Then $\mathsf{E}_\phi(\hom(G,H,\phi)) = t(G,H) \geq 0$, and $t(G,H) > 0$ if all weights are 1, so $\mathsf{E}_H\mathsf{E}_\phi(\hom(G,H,\phi)) > 0$. Hence there is a ϕ for which $\mathsf{E}_H(\hom(G,H,\phi) > 0$. But clearly $\mathsf{E}_H(\hom(G,H,\phi) = 0$ unless ϕ is an even homomorphism of G into itself. □

Let K_n denote the complete graph on the vertex set $[n]$, where $n \geq |V(G)|$.

Theorem 8.5. *If a graph G is positive, then there exists an even homomorphism $f: G \to K_n$ so that $\big|f(V(G))\big| \geq \frac{1}{2}\big|V(G)\big|$.*

We will prove this theorem in Section 8.4.

There are certain operations on graphs that preserve symmetry. Every such operation should also preserve positiveness. We are going to prove three results of this kind; such results are also useful in proving the conjecture for small graphs.

We need some basic properties of the homomorphism density function: Let G_1 and G_2 be two simple graphs, and let G_1G_2 denote their disjoint union. Then for every kernel W,

$$t(G_1G_2, W) = t(G_1, W)t(G_2, W). \tag{8.2}$$

For two looped-simple graphs G_1 and G_2, we denote by $G_1 \times G_2$ their *categorical product*, defined by

$$\begin{aligned} V(G_1 \times G_2) &= V(G_1) \times V(G_2), \\ E(G_1 \times G_2) &= \big\{\big((i_1,i_2),(j_1,j_2)\big) : \ (i_1,j_1) \in E(G_1), (i_2,j_2) \in E(G_2)\big\}. \end{aligned}$$

We note that if at least one of G_1 and G_2 is simple (has no loops) then so is the product. The quantity $t(G_1 \times G_2, W)$ cannot be expressed as simply as (8.2), but the following formula will be good enough for us. For a kernel W and looped-simple graph G, let us define the function $W^G: \ ([0,1]^V)^2 \to [0,1]$ by

$$W^G\big((x_1,\dots,x_k),(y_1,\dots,y_k)\big) = \prod_{(i,j)\in E(G)} W(x_i,y_j) \tag{8.3}$$

(every non-loop edge of G contributes two factors in this product). Then we have

$$t(G \times H, W) = t(G, W^H). \tag{8.4}$$

The following lemma implies that it is enough to prove the conjecture for connected graphs.

Lemma 8.6. *A graph G is positive if and only if every connected graph that occurs among the connected components of G an odd number of times is positive.*

Proof. The "if" part is obvious by (8.2). To prove the converse, let $G_1,\dots,G_m$ be the connected components of a positive graph G. We may assume that these connected components are different and they are non-positive, since omitting a positive component or two isomorphic components does not change positivity of G. We want to show that $m = 0$. Suppose that $m \geq 1$.

Claim 8.1. *We can choose kernels $W_1, \dots, W_m$ so that $t(G_i, W_i) < 0$ and $t(G_i, W_j) \neq t(G_j, W_j)$ for $i \neq j$.*

For every i there is a kernel W_i such that $t(G_i, W_i) < 0$, since G_i is not positive. Next we show that for every $i \neq j$ there is a kernel W_{ij} such that $t(G_i, W_{ij}) \neq t(G_j, W_{ij})$. If $|V(G_i)| \neq |V(G_j)|$ then the kernel $W_{ij} = \mathbb{1}(x, y \leq 1/2)$ does the job, so suppose that $|V(G_i)| = |V(G_j)|$. Then there is a simple graph H such that $\hom(G_i, H) \neq \hom(G_j, H)$, and hence we can choose $W_{ij} = W_H$.

Let $W_j' = W_j + \sum_{i \neq j} x_i W_{ij}$, then $t(G_i, W_j')$, $(i = 1, \dots, m)$ are different polynomials in the variables x_i, and hence their values are different for a generic choice of the x_i. If the x_i are chosen close to 0, then $t(G_j, W_j') < 0$, and hence we can replace W_j by W_j'. This proves the Claim.

Let W_0 denote the identically-1 kernel. For nonnegative integers $k_0, \dots, k_m$, construct a kernel $W_{k_0,\dots,k_m}$ by taking the direct sum of k_i copies of W_i. Then

$$t(G_1 \dots G_m, W_{k_0,\dots,k_m}) \overset{(8.2)}{=} \prod_{j=1}^{m} \Big(\sum_{i=0}^{m} k_i t(G_j, W_i) \Big).$$

We know that this expression is nonnegative for every choice of the k_i. Since the right hand side is homogeneous in $k_0, \dots, k_m$, it follows that

$$\prod_{j=1}^{m} \Big(1 + \sum_{i=1}^{m} x_i t(G_j, W_i) \Big) \geq 0 \tag{8.5}$$

for every $x_1, \dots, x_m \geq 0$. But the m linear forms $\ell_j(x) = 1 + \sum_{i=1}^{m} x_i t(G_j, W_i)$ are different by the choice of the W_i, and each of them vanishes on some point of the positive orthant since $t(G_j, W_j) < 0$. Hence there is a point $x \in \mathbb{R}_+^m$ where the first linear form vanishes but the other forms do not. In a small neighborhood of this point the product (8.5) changes sign, which is a contradiction. □

Proposition 8.7. *If G is a positive simple graph and H is any looped-simple graph, then $G \times H$ is positive.*

Proof. Immediate from (8.4). □

Let $G(r)$ be the graph obtained from G by replacing each node with r twins of it. Then $G(r) \cong K_r^\circ \times G$, where K_r° is the complete r-graph with a loop added at every node. Hence we get:

Corollary 8.8. *If G is positive, then so is $G(r)$ for every positive integer r.*

As a third result of this kind, we will show that the subgraph of a positive graph spanned by nodes with a given degree is also positive (Corollary 8.15). This proof, however, is more technical and is given in the next section. Unfortunately, these tools do not help us much for regular graphs G.

8.3 Subgraphs of positive graphs

In this section, let $G = (V, E)$ be a simple graph. For a measurable subset $\mathcal{F} \subseteq [0,1]^V$ and a bounded measurable weight function $\omega : [0,1] \to (0, \infty)$, we define

$$t(G, W, \omega, \mathcal{F}) = \int_{\mathcal{F}} \prod_{v \in V} \omega\big(p(v)\big) \prod_{e \in E} W\big(p(e)\big) dp. \tag{8.6}$$

With the measure μ with density function ω (i.e., $\mu(X) = \int_X \omega$), we can write this is

$$t(G, W, \omega, \mathcal{F}) = \int_{\mathcal{F}} \prod_{e \in E} W\big(p(e)\big) d\mu^V(p). \tag{8.7}$$

We say that G is *$\mathcal{F}$-positive* if for every kernel W and function ω as above, we have $t(G, W, \omega, \mathcal{F}) \geq 0$. It is easy to see that G is $[0,1]^V$-positive if and only if it is positive.

We say that $\mathcal{F}_1, \mathcal{F}_2 \subseteq [0,1]^V$ are *equivalent* if there exists a bijection $f : [0,1] \to [0,1]$ such that both f and f^{-1} are measurable, and $p \in \mathcal{F}_1 \Leftrightarrow f(p) \in \mathcal{F}_2$, where $f(p)(v) = f\big(p(v)\big)$.

Lemma 8.9. *If $\mathcal{F}_1$ and $\mathcal{F}_2$ are equivalent, then G is $\mathcal{F}_1$-positive if and only if it is $\mathcal{F}_2$-positive.*

Proof. Let f denote the bijection in the definition of the equivalence. For a kernel W and weight function ω, define the kernel $W^f(x,y) = W\big(f(x), f(y)\big)$, and weight function $\omega^f(x) = \omega\big(f(x)\big)$, and let μ and μ_f denote the measures defined by ω and ω^f, respectively. With this notation,

$$\begin{aligned} t(G, W^f, \omega_f, \mathcal{F}_2) &= \int_{\mathcal{F}_2} \prod_{e \in E} W^f\big(p(e)\big) d\mu_f^V(p) \\ &= \int_{\mathcal{F}_1} \prod_{e \in E} W\big(p(e)\big) d\mu^V(p) = t(G, W, \omega, \mathcal{F}_1). \end{aligned}$$

This shows that if G is $\mathcal{F}_2$-positive, then it is also $\mathcal{F}_1$-positive. The reverse implication follows similarly. □

For a nonnegative kernel $W : [0,1]^2 \to [0,1]$ (these are also called *graphons*), function $\omega : [0,1] \to [0,\infty)$, and $\mathcal{F} \subseteq [0,1]^V$, define

$$s = s(G, W, \omega, \mathcal{F}) = \sup_{p \in \mathcal{F}} \Big(\prod_{v \in V} \omega\big(p(v)\big) \cdot \prod_{e \in E} W\big(p(e)\big) \Big), \tag{8.8}$$

and

$$\mathcal{F}_{max} = \Big\{ p \in \mathcal{F} : \prod_{v \in V} \omega\big(p(v)\big) \cdot \prod_{e \in E} W\big(p(e)\big) = s \Big\}.$$

If the Lebesgue measure $\lambda(\mathcal{F}_{max}) > 0$, then we say that $\mathcal{F}_{max}$ is *emphasizable* from $\mathcal{F}$, and (W, α) emphasizes it.

Lemma 8.10. *If G is $\mathcal{F}_1$-positive and $\mathcal{F}_2$ is emphasizable from $\mathcal{F}_1$, then G is $\mathcal{F}_2$-positive.*

Proof. Suppose that (U, τ) emphasizes $\mathcal{F}_2$ from $\mathcal{F}_1$, and let $s = s(G, U, \tau, \mathcal{F}_1)$. Assume that G is not $\mathcal{F}_2$-positive, then there exists a kernel W and a weight function ω with $t(G, W, \omega, \mathcal{F}_2) < 0$. Consider the kernel $W_n = U^n W$ and weight function $\omega_n = s^{-n/|V|} \tau^n \omega$. Then

$$\prod_{v \in V} \omega_n\big(p(v)\big) \cdot \prod_{e \in E} W_n\big(p(e)\big) = \Big(\prod_{v \in V} \omega\big(p(v)\big) \cdot \prod_{e \in E} W\big(p(e)\big) \Big) \cdot a(p)^n,$$

where

$$a(p) = \frac{1}{s} \prod_{v \in V} \tau\big(p(v)\big) \cdot \prod_{e \in E} U\big(p(e)\big) \begin{cases} = 1 & \text{if } p \in \mathcal{F}_2, \\ < 1 & \text{otherwise.} \end{cases}$$

Thus

$$t(G,W_n,\omega_n,\mathcal{F}_1) = \int\limits_{\mathcal{F}_1} \prod_{v\in V} \omega_n\big(p(v)\big)\cdot \prod_{e\in E} W_n\big(p(e)\big)\,dp$$
$$\to \int\limits_{\mathcal{F}_2} \prod_{v\in V} \omega\big(p(v)\big)\cdot \prod_{e\in E} W\big(p(e)\big)\,dp = t(G,W,\omega,\mathcal{F}_2) < 0,$$

which implies that G is not $\mathcal{F}_1$-positive. □

For a partition $\mathcal{P}$ of $[0,1]$ into a finite number of sets with positive measure and function $f: V \to \mathcal{P}$, we call the box $\mathcal{F}(f) = \{p \in [0,1]^V : \ p(v) \in f(v)\ \forall v \in V\}$ a *partition-box*. Equivalently, a partition-box is a product set $\prod_{v\in V} S_v$, where the sets $S_v \subseteq [0,1]$ are measurable, and either $S_u \cap S_v = \emptyset$ or $S_u = S_v$ for all $u, v \in V$.

Lemma 8.11. *If $\mathcal{F}_1 \supseteq \mathcal{F}_2$ are partition-boxes, and G is $\mathcal{F}_2$-positive, then it is $\mathcal{F}_1$-positive.*

Proof. Let $\mathcal{F}_i$ be a product of classes of partition $\mathcal{P}_i$; we may assume that $\mathcal{P}_2$ refines $\mathcal{P}_1$. For $P \in \mathcal{P}_2$, let $\overline{P}$ denote the class of $\mathcal{P}_1$ containing P. We may assume that every partition class of $\mathcal{P}_1$ and $\mathcal{P}_2$ is an interval.

Consider any kernel W and any weight function ω. Let $\varphi: [0,1] \to [0,1]$ be the function that maps every $P \in \mathcal{P}_2$ onto $\overline{P}$ in a monotone and affine way. The map φ is measure-preserving, because for each $A \subseteq Q \in \mathcal{P}_1$,

$$\lambda\big(\varphi^{-1}(A)\big) = \sum_{\substack{P\in\mathcal{P}_2\\ P\subseteq Q}} \lambda\big(\varphi^{-1}(A)\cap P\big) = \sum_{\substack{P\in\mathcal{P}_2\\ P\subseteq Q}} \lambda(A)\frac{\lambda(P)}{\lambda(Q)} = \lambda(A). \tag{8.9}$$

Applying φ coordinate-by-coordinate, we get a measure preserving map $\psi: [0,1]^V \to [0,1]^V$. Then $\psi' = \psi|_{\mathcal{F}_2}$ is an affine bijection from $\mathcal{F}_2$ onto $\mathcal{F}_1$, and clearly $\det(\psi') > 0$. Hence

$$t(G,W^\varphi,\omega^\varphi,\mathcal{F}_2) \overset{(8.1)}{=} \int\limits_{\mathcal{F}_2} \prod_{v\in V} \omega^\varphi\big(p(v)\big)\cdot \prod_{e\in E} W^\varphi\big(p(e)\big)dp$$
$$= \det(\psi')\cdot \int\limits_{\mathcal{F}_1} \prod_{v\in V} \omega\big(p(v)\big)\cdot \prod_{e\in E} W\big(p(e)\big)dp$$
$$\overset{(8.1)}{=} \det(\psi')\cdot t(G,W,\omega,\mathcal{F}_1).$$

Since G is $\mathcal{F}_2$-positive, the left hand side is positive, and hence $t(G,W,\omega,\mathcal{F}_1) \geq 0$, proving that G is $\mathcal{F}_1$-positive. □

Lemma 8.12. *Suppose that $\mathcal{F}_1$ is a partition-box defined by a partition $\mathcal{P}$ and function f_1. Let $Q \in \mathcal{P}$ and let U be the union of an arbitrary set of classes of $\mathcal{P}$. Let θ be a positive number but not an integer. Split Q into two parts with positive measure, Q^+ and Q^-. Let $\deg(v,U)$ denote the number of neighbors u of v with $f_1(u) \subseteq U$. Define*

$$f_2(v) = \begin{cases} f_1(v) & \text{if } f_1(v) \neq Q,\\ Q^+ & \text{if } f_1(v) = Q \text{ and } \deg(v,U) > d,\\ Q^- & \text{if } f_1(v) = Q \text{ and } \deg(v,U) < d,\end{cases}$$

and let $\mathcal{F}_2$ be the corresponding partition-box. Then there exists a pair (W,ω) emphasizing $\mathcal{F}_2$ from $\mathcal{F}_1$.

Proof. Clearly, $\lambda(\mathcal{F}_2) > 0$. First, suppose that $Q \not\subseteq U$. Let W be 2 in $Q^+ \times U$ and in $U \times Q^+$, and 1 everywhere else. Let $\omega(x)$ be 2^{-d} if $x \in Q^+$ and 1 otherwise. It is easy to see that the weight of a $p \in \mathcal{F}_1$ is 2^a, where $a = \sum_{v \in p^{-1}(Q^+)} \big(\deg(v, U) - d\big)$. This expression is maximal if and only if $p \in \mathcal{F}_2$. The case when $Q \subset U$ is similar. □

We can use Lemma 8.12 iteratively: we start with the indiscrete partition, and refine it so that G remains positive relative to partition-boxes of these partitions. This is essentially the 1-diemsnional Weisfeiler–Lehman algorithm. There is a non-iterative description of the resulting partition, and this is what we are going to describe next.

The *walk-tree* of a rooted graph (G, v) is the following infinite rooted tree $R(G, v)$. Its nodes are all finite walks starting from v, its root is the 0-length walk, and the parent of any other walk is obtained by deleting its last node. Let $\mathcal{R}$ be the partition of V in which two nodes $u, v \in V$ belong to the same class if and only if $R(G, u) \cong R(G, v)$. A function $f : V \to \mathcal{P}$ is a *walk-tree function* if $\mathcal{P}$ is a measurable partition of $[0, 1]$, and f is constant on every class of $\mathcal{R}$.

Proposition 8.13. *If a graph G is positive, then for every kernel W, weight function ω, and partition-box $\mathcal{F}(f)$ defined by a walk-tree function f, we have $t(G, W, \omega, \mathcal{F}) \geq 0$.*

Proof. Let the k-neighborhood of r in $R(G, r)$ be denoted by $R_k(G, r)$. We say that a function $f : V \to \mathcal{P}$ is a k-walk-tree function if $R_k(G, u) = R_k(G, v)$ whenever $f(u) = f(v)$ $(u, v \in V)$. Every walk-tree function is a k-walk-tree function with a sufficiently large k. Thus it suffices to prove the proposition for all k-walk-tree functions f.

We prove this by induction. If $k = 0$, then the condition is the same as the assertion. Now, let us assume that the statement is true for a k. Using Lemmas 8.11 and 8.12, we separate each class according to the number of neighbors in the different other classes. This way we divide the classes according to the $(k + 1)$-walk-trees. □

Corollary 8.14. *If G is positive, then the subgraph spanned by the preimage of an arbitrary set under a walk-tree function is also positive.*

Proof. Suppose that the subgraph is negative with some W. Let us extend (and then renormalize) the ground set $[0, 1]$ with one more class for the other nodes of G, and set $W = 1$ at the extension of the domain of W. This way we get the same negative homomorphism number, which remains negative after renormalization. □

Corollary 8.15. *If G is positive, then for each k, the subgraph of G spanned by all nodes with degree k is positive as well.* □

Corollary 8.16. *For each odd k, the number of nodes of G with degree k must be even.*

Proof. Otherwise, consider the partition-box $\mathcal{F}$ that separates the vertices of G with degree d to class $A = [0, 1/2]$ and the other vertices to $\bar{A} = (1/2, 1]$. Consider the kernel W which is -1 between A and $\bar{A}$ and 1 in the other two cells. Then for each map $p \in [0, 1]^V$, the total degree of the nodes mapped into class A is odd, so there is an odd number of edges between A and $\bar{A}$. So the weight of p is -1, therefore $t(G, W, 1, \mathcal{F}) = -\lambda(\mathcal{F}) < 0$. □

Corollary 8.17. *Conjecture 8.1 is true for trees.*

Proof. >From the walk-tree of a vertex v of the tree G, we can easily decode the rooted tree G. Let us make the walk-tree decomposition as in Proposition 8.13. We call a vertex *central* if it cuts G into components with at most $|V|/2$ nodes. There can be either one central node or two neighboring central nodes of G. If there are two of them, then their walk-trees are different

from the walk-trees of every other nodes, but these two points span one edge, which is not positive, therefore Lemma 8.14 implies that neither is G. If there is only one central node, then consider the walk-trees of its neighbors. If there is an even number of each kind, then G is symmetric. Otherwise we can find two classes with an odd number of edges between them, which is not positive. □

8.4 Homomorphic images of positive graphs

The main goal of this section is to prove Theorem 8.5. In what follows, let n be a large integer. For a homomorphism $f: G \to K_n$, we call an edge $e \in E(K_n)$ *f-odd* if $\big|f^{-1}(e)\big|$ is odd. We call a vertex $v \in V(K_n)$ *f-odd* if there exists an f-odd edge incident with v. Let $E_{\text{odd}}(f)$ and $V_{\text{odd}}(f)$ denote the set of f-odd edges and nodes of K_n, respectively, and define

$$r(f) = \big|V(G)\big| - \big|f\big(V(G)\big)\big| + \frac{1}{2}|V_{\text{odd}}(f)|. \tag{8.10}$$

Lemma 8.18. *Let $G_i = (V_i, E_i)$ $(i = 1, 2)$ be two graphs, let $f: G_1G_2 \to K_n$, and let $f_i: G_i \to K_n$ denote the restriction of f to V_i. Then $r(f) \geq r(f_1) + r(f_2)$.*

Proof. Clearly $|V(G)| = |V_1| + |V_2|$ and $|V(f(G))| = |f(V_1)| + |f(V_2)| - |f(V_1) \cap f(V_2)|$. Furthermore, $E_{\text{odd}}(f) = E_{\text{odd}}(f_1) \triangle E_{\text{odd}}(f_2)$, which implies that $V_{\text{odd}}(f) \supseteq V_{\text{odd}}(f_1) \triangle V_{\text{odd}}(f_2)$. Hence

$$\begin{aligned} |V_{\text{odd}}(f)| &\geq |V_{\text{odd}}(f_1)| + |V_{\text{odd}}(f_2)| - 2|V_{\text{odd}}(f_1) \cap V_{\text{odd}}(f_2)| \\ &\geq |V_{\text{odd}}(f_1)| + |V_{\text{odd}}(f_2)| - 2|f(V_1) \cap f(V_2)|. \end{aligned}$$

Substituting these expressions in (8.10), the lemma follows. □

Let G^k denote the disjoint union of k copies of a graph G. This lemma implies that if $f: G^k \to K_n$ is any homomorphism and $f_i: G \to K_n$ denotes the restriction of f to the i-th copy of G, then

$$r(f) \geq \sum_{i=1}^{k} r(f_i). \tag{8.11}$$

We define two parameters of a graph G:

$$p(G) = \min\Big\{|V(G)| - |f(V(G))| \;\Big|\; f: G \to K_n \text{ is even}\Big\} \tag{8.12}$$

and

$$\bar{r}(G) = \min\big\{r(f)\big|f: G \to K_n\big\}. \tag{8.13}$$

Since $p(G) = \min\big\{r(f)\big|f: G \to K_n \text{ is even}\big\}$, it follows that

$$p(G) \geq \bar{r}(G). \tag{8.14}$$

Furthermore, considering any injective $f: G \to K_n$, we see that

$$\bar{r}(G) \leq r(f) = \big|V(G)\big| - \big|f\big(V(G)\big)\big| + \frac{1}{2}\big|f\big(V(G)\big)\big| = \frac{1}{2}\big|V(G)\big|. \tag{8.15}$$

Lemma 8.19.

$$\bar{r}(G^k) = k\bar{r}(G). \tag{8.16}$$

Proof. For one direction, take an $f:\ G^k \to K_n$ with $r(f) = \bar{r}(G^k)$. Then

$$\bar{r}(G^k) = r(f) \overset{(8.11)}{\geq} \sum_{i=1}^{k} r(f_i) \overset{(8.13)}{\geq} \sum_{i=1}^{k} \bar{r}(G) = k \cdot \bar{r}(G).$$

For the other direction, let us choose each f_i so that $r(f_i) = \bar{r}(G)$ and the images $f_i(G)$ are pairwise disjoint. Then

$$\bar{r}(G^k) \overset{(8.13)}{\leq} r(f) = \sum_{i=1}^{k} r(f_i) = \sum_{i=1}^{k} \bar{r}(G) = k \cdot \bar{r}(G). \qquad \square$$

Lemma 8.20.

$$p(G^2) = \bar{r}(G^2). \tag{8.17}$$

Proof. We already know by (8.14) that $p(G^2) \geq \bar{r}(G^2)$. For the other direction, we define $f:\ G^2 \to K_n$ as follows. We choose f_1 so that $r(f_1) = \bar{r}(G)$. Consider all points $v_1, v_2, ..., v_l$ in $f\big(V(G)\big)$ which are not f_1-odd. Let us choose pairwise different nodes $v_1', v_2', ..., v_l'$ disjointly from $f\big(V(G)\big)$. Now we choose f_2 so that for each $x \in V(G)$, if $f_1(x)$ is an f_1-odd point, then $f_2(x) = f_1(x)$, and if $f_1(x) = v_i$, then $f_2(i) = v_i'$.

If an edge $e \in E(K_n)$ is incident to a v_i, then $\big|f_1^{-1}(e)\big|$ is even and $f_2^{-1}(e) = \emptyset$. If e is incident to a v_i', then $\big|f_2^{-1}(e)\big|$ is even and $f_1^{-1}(e) = \emptyset$. If e is not incident to any v_i or v_i', then $\big|f_1^{-1}(e)\big| = \big|f_2^{-1}(e)\big|$. Therefore f is even. Thus,

$$\begin{aligned} p(G^2) &\overset{(8.12)}{\leq} r(f) \overset{(8.10)}{=} \big|V(G^2)\big| - \big|f(V(G^2))\big| \\ &= 2\big|V(G)\big| - \Big|f_1\big(V(G)\big)\Big| - \Big|f_2\big(V(G)\big)\Big| + \Big|f_1\big(V(G)\big) \cap f_2\big(V(G)\big)\Big| \\ &= 2\big|V(G)\big| - 2\Big|f_1\big(V(G)\big)\Big| + o\Big(f_1\big(V(G)\big)\Big) \overset{(8.10)}{=} 2r(f_1) = 2\bar{r}(G) \overset{(8.16)}{=} \bar{r}(G^2). \end{aligned}$$ $\square$

Let K_n^w denote K_n equipped with an edge-weighting $w:\ E(K_n) \to \{-1, 1\}$. Let the stochastic variable $K_n^{\pm 1}$ denote K_n^w with a uniform random w.

Lemma 8.21. *For a fix graph G, and $n \to \infty$,*

$$\mathsf{E}\big(t(G, K_n^{\pm 1})\big) = \Theta\big(n^{-p(G)}\big).$$

Proof. If an edge e is f-odd, then changing the weight on e changes the sign of the homomorphism, therefore $\mathsf{E}_w\big(\hom(G, K_n^w, f)\big) = 0$. On the other hand, if f is even, then for all w, $\hom(G, K_n^w, f) = 1$. Therefore, taking a uniformly random homomorphism $f:\ G \to K_n$,

$$\begin{aligned} \mathsf{E}\big(t(G, K_n^{\pm 1})\big) &= \mathsf{E}_w\Big(\mathsf{E}_f\big(\hom(G, K_n^w, f)\big)\Big) = \mathsf{E}_f\Big(\mathsf{E}_w\big(\hom(G, K_n^w, f)\big)\Big) \\ &= \mathsf{P}(f \text{ is even}). \end{aligned}$$

Clearly,

$$\mathsf{P}(f \text{ is even}) \leq \mathsf{P}\Big(\big|V(G)\big| - \big|V(f(G))\big| \geq p(G)\Big) = O(n^{-p(G)}).$$

On the other hand, consider an even homomorphism $g:\ G \to K_n$ with $r(g) = p(G)$. We say that $f, g:\ G \to K_n$ are isomorphic if there exists a permutation σ on $V(K_n)$ that $\forall x \in V(G):\ f(x) = \sigma\big(g(x)\big)$. There are $\binom{n}{|g(V(G))|}$ different functions isomorphic with g. Therefore,

$$\begin{aligned} \mathsf{P}(f \text{ is even}) &\geq \mathsf{P}(f \text{ is isomorphic with } g) = \binom{n}{|g(V(G))|} \Big/ n^{|V(G)|} \\ &= \Omega(n^{-p(G)}). \end{aligned}$$

$\square$

Now let us turn to the proof of Theorem 8.5. Assume that G is positive, then the random variable $X = t(G, K_n^{\pm 1})$ is nonnegative. Applying Hölder's inequality to $X^{1/2}$ and $X^{3/2}$ with $p = q = 2$, we get that

$$\mathsf{E}(X) \cdot \mathsf{E}(X^3) \geq \mathsf{E}(X^2)^2. \tag{8.18}$$

Here

$$\mathsf{E}(X^k) = \mathsf{E}\big(t(G, K_n^{\pm 1})^k\big) \overset{(8.2)}{=} \mathsf{E}\big(t(G^k, K_n^{\pm 1})\big) = \Theta\big(n^{-p(G^k)}\big),$$

so (8.18) shows that $n^{-p(G)} \cdot n^{-p(G^3)} = \Omega\big(n^{-2p(G^2)}\big)$, thus $p(G) + p(G^3) \leq 2p(G^2)$. Hence

$$4\bar{r}(G) \overset{(8.16)}{=} \bar{r}(G) + \bar{r}(G^3) \leq p(G) + p(G^3) \leq 2p(G^2) \overset{(8.17)}{=} 2\bar{r}(G^2) \overset{(8.16)}{=} 4\bar{r}(G). \tag{8.19}$$

All expressions in (8.19) must be equal, therefore $\bar{r}(G) = p(G)$.

Finally, for an even $f : G \to K_n$ with $\big|V(G)\big| - \big|f(V(G))\big| = p(G)$, we have

$$\frac{1}{2}\big|V(G)\big| \overset{(8.15)}{\geq} \bar{r}(G) = p(G) = \big|V(G)\big| - \big|f(V(G))\big|,$$

therefore $\big|f\big(V(G)\big)\big| \geq \frac{1}{2}\big|V(G)\big|$.

8.5 Computational results

We checked the conjecture for all graphs on at most 9 vertices using the previous results and a computer program. Starting from the list of nonisomorphic graphs, we filtered out those who violated one of our conditions for being a minimal counterexample. In particular we performed the following tests:

1. Check whether the graph is symmetric, by exhaustive search enumerating all possible involutions of the vertices.

2. Calculate the number of homomorphisms into graphs represented by 1×1, 2×2 or 3×3 matrices of small integers. (Checking 1×1 matrices is just the same as checking whether or not the number of edges is even.) If we get a negative homomorphism count, the graph is negative and therefore it is not a counterexample.

3. Calculate the number of homomorphisms into graphs represented by symbolic 3×3 and 4×4 matrices and perform local minimization on the resulting polynomial from randomly chosen points. Once we reach a negative value, we can conclude that the graph is negative.

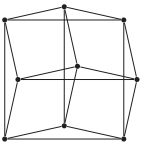
4. Partition the vertices of the graph in such a way that two vertices belong to the same class if and only if they produce the same walk-tree (1-dimensional Weisfeiler–Lehman Algorithm). Check for all proper subsets of the set of classes whether their union spans an asymmetric subgraph. If we find such a subgraph, the graph is not a minimal counterexample: either the subgraph is not positive and by Corollary 8.14 the original graph is not positive either, or the subgraph is positive, and therefore we have a smaller counterexample.

5. Consider only those homomorphisms which map all vertices in the ith class of the partition into vertices $3i+1$, $3i+2$ and $3i+3$ of the target graph represented by a symbolic matrix. If we get a negative homomorphism count, the graph is negative by Proposition 8.13. (In this case we work with a $3k \times 3k$ matrix where k denotes the number of classes of the

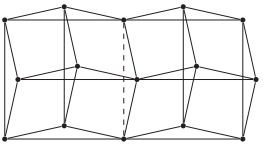
walk-tree partition, but the resulting polynomial still has a manageable size because we only count a small subset of homomorphisms. Note that if one of the classes consists of a single vertex, we only need one corresponding vertex in the target graph.)

The tests were performed in such an order that the faster and more efficient ones were run first, restricting the later ones to the set of remaining graphs. For example, in step 4, we start with checking whether any of the classes spans an odd number of edges, or whether the number of edges between any two classes is odd. We used the `SAGE` computer-algebra system for our calculations and rewritten the speed-critical parts in `C` using `nauty` for isomorphism checking, `mpfi` for interval arithmetics and Jean-Sébastien Roy's `tnc` package for nonlinear optimization.

Our automated tests left only one graph on 9 vertices as a possible minimal counterexample, the graph on left:

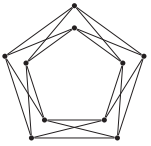


G_1

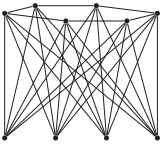


H

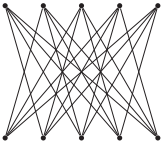
The non-positivity of this graph was checked manually by counting the number of homomorphisms into the graph on the right (where the dashed edge has weight -1 and all other edges have weight 1). This leaves only the following three of the 12 293 435 graphs on at most 10 vertices as candidates for a minimal counterexample:



G_2



G_3



G_4

Note that all three graphs are regular, as is the case for all remaining graphs on 11 vertices. We have found step 5 of the algorithm quite effective at excluding graphs with nontrivial walk-tree partitions.

Bibliography

[1] David Aldous and Russell Lyons. Processes on unimodular random networks. *Electron. J. Probab.*, 12:no. 54, 1454–1508, 2007.

[2] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of algorithms*, 7(4):567–583, 1986.

[3] D. Angluin. Local and global properties in networks of processors. In *Proceedings of the twelfth annual ACM symposium on Theory of computing*, pages 82–93. ACM, 1980.

[4] M. Bayati, D. Gamarnik, and P. Tetali. Combinatorial approach to the interpolation method and scaling limits in sparse random graphs. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 105–114. ACM, 2010.

[5] Itai Benjamini and Oded Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:no. 23, 13 pp. (electronic), 2001.

[6] Robert Berger. The undecidability of the domino problem. *Mem. Amer. Math. Soc. No.*, 66:72, 1966.

[7] B. Bollobás. The independence ratio of regular graphs. *Proc. Amer. Math. Soc.*, 83(2):433–436, 1981.

[8] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, B. Szegedy, and K. Vesztergombi. Graph limits and parameter testing. In *Annual ACM Symposium on Theory of Computing: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, volume 21, pages 261–270, 2006.

[9] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, and K. Vesztergombi. Counting graph homomorphisms. *Topics in discrete mathematics*, pages 315–371, 2006.

[10] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós, and K. Vesztergombi. Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851, 2008.

[11] L. Bowen. Couplings of uniform spanning forests. *Proceedings of the American Mathematical Society*, pages 2151–2158, 2004.

[12] V. K. Bulitko. On the problem of the finiteness of a graph with given vertex neighborhoods. In *General systems theory (Russian)*, pages 76–83. Akad. Nauk Ukrain. SSR Inst. Kibernet., Kiev, 1972.

[13] V. K. Bulitko. Graphs with prescribed environments of the vertices. *Trudy Mat. Inst. Steklov.*, 133:78–94, 274, 1973. Mathematical logic, theory of algorithms and theory of sets (dedicated to P. S. Novikov on the occasion of his seventieth birthday).

[14] O.A. Camarena, E. Csóka, T. Hubai, G. Lippner, and L. Lovász. Positive graphs. *arXiv preprint arXiv:1205.6510*, 2012.

[15] C.T. Conley, A.S. Kechris, and R.D. Tucker-Drob. Ultraproducts of measure preserving actions and graph combinatorics. *Ergodic Theory and Dynamical Systems*, 1(1):1–41, 2011.

[16] E. Csóka. Maximum flow is approximable by deterministic constant-time algorithm in sparse networks. *arXiv preprint arXiv:1005.0513*, 2010.

[17] E. Csóka. Local algorithms with public randomisation on sparse graphs. *arXiv preprint arXiv:1202.1565*, 2012.

[18] E. Csóka. An undecidability result on limits of sparse graphs. *The Electronic Journal of Combinatorics*, 19(2):P21, 2012.

[19] E. Csóka and G. Lippner. Invariant random matchings in cayley graphs. *arXiv preprint arXiv:1211.2374*, 2012.

[20] G. Elek. Note on limits of finite graphs. *Combinatorica*, 27(4):503–507, 2007.

[21] G. Elek. On the limit of large girth graph sequences. *Combinatorica*, 30(5):553–563, 2010.

[22] G. Elek. Parameter testing in bounded degree graphs of subexponential growth. *Random Structures & Algorithms*, 37(2):248–270, 2010.

[23] G. Elek. Samplings and observables. convergence and limits of metric measure spaces. *arXiv preprint arXiv:1205.6936*, 2012.

[24] G. Elek and G. Lippner. Borel oracles. an analytical approach to constant-time algorithms. In *Proc. Amer. Math. Soc*, volume 138, pages 2939–2947, 2010.

[25] L.R. Ford and D.R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8(3):399–404, 1956.

[26] D. Gaboriau and R. Lyons. A measurable-group-theoretic solution to von neumann's problem. *Inventiones mathematicae*, 177(3):533–540, 2009.

[27] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 339–348. IEEE, 1996.

[28] O. Goldreich and D. Ron. Property testing in bounded degree graphs. *Algorithmica*, 32(2):302–343, 2002.

[29] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Birkhäuser Boston, 2006.

[30] H. Hatami. Graph norms and sidorenko's conjecture. *Israel Journal of Mathematics*, 175(1):125–150, 2010.

[31] H. Hatami, L. Lovász, and B. Szegedy. Limits of local-global convergent graph sequences. *arXiv preprint arXiv:1205.4356*, 2012.

[32] Hamed Hatami and Serguei Norine. Undecidability of linear inequalities in graph homomorphism densities. *J. Amer. Math. Soc.*, 24(2):547–565, 2011.

[33] J. Hladkỳ. Induced bipartite subgraphs in a random cubic graph. 2007.

[34] C. Hoppen. Properties with graphs of large girth. *PhD Thesis, University of Waterloo*, 2008.

[35] C. Hoppen, Y. Kohayakawa, C.G. Moreira, B. Rath, and R. Menezes Sampaio. Limits of permutation sequences. *Journal of Combinatorial Theory, Series B*, 2012.

[36] A. Israeli and A. Itai. A fast and simple randomized parallel algorithm for maximal matching. *Information Processing Letters*, 22(2):77–80, 1986.

[37] S. Janson. Poset limits and exchangeable random posets. *Combinatorica*, pages 1–35, 2009.

[38] M. Jerrum and U. Vazirani. A mildly exponential approximation algorithm for the permanent. *Algorithmica*, 16(4):392–401, 1996.

[39] F. Kardoš, D. Král, and J. Volec. Fractional colorings of cubic graphs with large girth. *Preprint*, 2010. arXiv:1010.3415v1.

[40] M. Laczkovich. Equidecomposability and discrepancy; a solution of tarski's circle squaring problem. *J. Reine Angew. Math*, 404:77–117, 1990.

[41] J. Lauer and N. Wormald. Large independent sets in regular graphs of large girth. *J. Combin. Theory Ser. B*, 97(6):999–1009, 2007.

[42] N. Linial. Locality in distributed graph algorithms. *SIAM Journal on Computing*, 21:193, 1992.

[43] Yang-Yu Liu, Endre Csóka, Haijun Zhou, and Márton Pósfai. Core percolation on complex networks. *Phys. Rev. Lett.*, 109:205703, Nov 2012.

[44] László Lovász and Balázs Szegedy. Limits of dense graph sequences. *J. Combin. Theory Ser. B*, 96(6):933–957, 2006.

[45] L. Lovász. Very large graphs. *Arxiv preprint arXiv:0902.0132*, 2009.

[46] M. Luby. A simple parallel algorithm for the maximal independent set problem. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 1–10. ACM, 1985.

[47] R. Lyons and F. Nazarov. Perfect matchings as iid factors on non-amenable groups. *European Journal of Combinatorics*, 32(7):1115–1125, 2011.

[48] B. D. McKay. Independent sets in regular graphs of high girth. *Ars Combin.*, 23A:179–185, 1987.

[49] M. Naor and L. Stockmeyer. What can be computed locally? In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 184–193. ACM, 1993.

[50] H.N. Nguyen and K. Onak. Constant-time approximation algorithms via local improvements. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 327–336. IEEE, 2008.

[51] J. B. Shearer. A note on the independence number of triangle-free graphs. *Discrete Math.*, 46(1):83–87, 1983.

[52] J. B. Shearer. A note on the independence number of triangle-free graphs, II. *J. Combin. Theory Ser. B*, 53(2):300–307, 1991.

[53] J. Suomela. Survey of local algorithms, 2009.

[54] B. Szegedy. Gowers norms, regularization and limits of functions on abelian groups. *arXiv preprint arXiv:1010.6211*, 2010.

[55] H. Wang, American Telephone, and Telegraph Company. *Proving theorems by pattern recognition-II.* American Telephone and Telegraph Company, 1961.

References for Chapter 7

[56] S. H. Strogatz, Nature **410**, 268 (2001).

[57] A.-L. Barabási, Nature Physics **8**, 14 (2011).

[58] A. Vespignani, Nature Physics **8**, 32 (2011).

[59] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[60] M. E. J. Newman, SIAM Review **45**, 167 (2003).

[61] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. **80**, 1275 (2008).

[62] P. Erdős and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[63] B. Bollobás, *Random Graphs* (Cambridge University Press, Cambridge, 2001).

[64] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. Lett. **85**, 5468 (2000).

[65] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[66] B. Pittel, J. Spencer, and N. Wormald, Journal of Combinatorial Theory **67**, 111 (1996).

[67] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Phys. Rev. Lett. **96**, 040601 (2006).

[68] A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes, Phys. Rev. E **73**, 056101 (2006), pRE.

[69] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Nature **435**, 814 (2005), 10.1038/nature03607.

[70] I. Derényi, G. Palla, and T. Vicsek, Phys. Rev. Lett. **94**, 3 (2005).

[71] D. Achlioptas, R. M. D'Souza, and J. Spencer, Science (New York, N.Y.) **323**, 1453 (2009).

[72] R. da Costa, S. Dorogovtsev, a. Goltsev, and J. Mendes, Phys. Rev. Lett. **105**, 2 (2010).

[73] O. Riordan and L. Warnke, Science **333**, 322 (2011).

[74] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, Phys. Rev. Lett. **85**, 4626 (2000).

[75] R. Albert, H. Jeong, and A.-L. Barabási, Nature **406**, 378 (2000).

[76] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, Nature **464**, 1025 (2010).

[77] R. Parshani, S. Buldyrev, and S. Havlin, Phys. Rev. Lett. **105**, 23 (2010).

[78] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin, Nature Physics **8**, 40 (2011a).

[79] J. Gao, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **107**, 195701 (2011b), pRL.

[80] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[81] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, Nature Physics **6**, 888 (2010).

[82] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, Nature **473**, 167 (2011).

[83] T. Nepusz and T. Vicsek, Nat Phys **advance online publication** (2012), 10.1038/nphys2327.

[84] M. Pósfai, Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, arXiv:1203.5161 (2012).

[85] T. Jia, Y.-Y. Liu, M. Pósfai, J.-J. Slotine, and A.-L. Barabási, *Control capability of complex networks*, unpublished.

[86] M. Bauer and O. Golinelli, Phys. Rev. Lett. **86**, 2621 (2001a).

[87] M. Bauer and O. Golinelli, Eur. Phys. J. B **24**, 339 (2001b).

[88] R. M. Karp and M. Sipser, Proc. 22nd Annual IEEE Symp. on Foundations of Computer Science pp. 364–375 (1981).

[89] H. Zhou and Z. Ou-Yang, *Maximum matching on random graphs*, *arXiv:cond-mat/0309348v1* (2003).

[90] L. Zdeborová and M. Mézard, J. Stat. Mech. **05**, P05003 (2006).

[91] M. Weigt and A. K. Hartmann, Phys. Rev. Lett. **84**, 6118 (2000).

[92] H. Zhou, Eur. Phys. J. B **32**, 265 (2003).

[93] A. K. Hartmann and M. Weigt, J. Phys. A **36**, 11069 (2003).

[94] H. Zhou, *Spin glass and message-passing* (in preparation, 2012).

[95] A. K. Hartmann, A. Mann, and W. Radenbach, Journal of Physics: Conference Series **95**, 012011 (2008).

[96] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (New York: W.H. Freeman, 1979).

[97] W. Barthel and A. K. Hartmann, Phys. Rev. E **70**, 066120 (2004).

[98] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).

[99] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).

[100] M. Catanzaro and R. Pastor-Satorras, Eur. Phys. J. B **44**, 241 (2005).

[101] J.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim, Eur. Phys. J. B **49**, 231 (2006).

[102] G. Parisi and T. Rizzo, Phys. Rev. E **78**, 022101 (2008), pRE.

[103] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).

[104] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).

[105] G. Bianconi, N. Gulbahce, and A. E. Motter, Phys. Rev. Lett. **100**, 118701 (2008).

[106] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[107] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 205 (2004).

[108] F. Chung and L. Lu, Annals of Combinatorics **6**, 125 (2002).

Summary

Very large graphs appear in biological systems, e.g. the brain; in physics, e.g. the graph of the bonds between the molecules of a solid; and so are the internet, the traffic system, the electrical grid, social networks, etc. Most of these graphs are not only huge but it is hopeless to get to know them precisely. However, we still have a chance to get to know some important properties and parameters of them.

In the model by Goldreich and Ron [28], we deal with bounded-degree graphs, and a (constant-size) *sample* from a graph G means the following. For some constants r and n, we take the (radius) r neighborhoods of n uniform random nodes of G. A graph parameter is *estimable* if for each $\varepsilon > 0$, there exists a function called *estimator* such that for all graphs G, the following holds. If the estimator receives a random sample from G as input, then this outputs a value with an error at most ε from the parameter value of G, in expectation. Some simple examples for these parameters are the expansion, the proportion of the nodes in the largest independent set, or dominating set, or matching.

A central question in this theory is what distributions of r-neighborhoods can be obtained from graphs. This question is related even to group-theoretic problems.

Local algorithm means a mapping from the isomorphism types of all rooted r-neighborhoods. For example, making an independent set on a graph by local algorithm means that whether each node is in the set depends only on its r-neighborhood. As an extension, we can assign independent random seeds to the nodes, and the output at each node can depend also on the random seeds assigned to the vertices in the r-neighborhood. There are further extensions, such as using a global random seed, or using some information about the isomorphism type of the entire graph.

For typical problems, we expect from local algorithms approximate solutions only. For example, we say that we can find an almost maximum independent set if for each $\varepsilon > 0$, there exists a local algorithm that for each graph G, outputs an independent set, and the expected size of this set is at most εn less than the size of the maximum independent set in G.

Local algorithms are strongly related to parameter estimation. For example, if we have a local algorithm which provides an almost maximum matching, then the relative size of the maximum matching is estimable.

In Chapter 2, we show a local algorithm finding an almost maximum flow and an almost minimum cut. Then we show important applications of this about neighborhood distributions of graphs. In Chapter 3, we show that, for local algorithms, sending the isomorphism type of the whole graph or sending only a global random seed are equally strong tools. In Chapter 4, we show the algorithmic undecidability of a specific class of questions about the possible neighborhood distributions of graphs. Chapter 5 is about the strength of local algorithms on an interesting specific problem, namely, constructing a large independent set on 3-regular large-girth graphs. In Chapter 6, we show a perfect matching on all nonamenable Cayley-graphs by a kind of limit of random local algorithms. Chapter 7 is about an application of the theory, as an example of how physicists research this topic, and how it is related to the mathematical approach. Finally, Chapter 8 is an example of how such theories can be useful for other areas of mathematics.

Chapters 2, 3 and 4 are results of the author, based on the papers [16], [17] and [18], respectively. Chapter 5 is a joint work with Gerencsér, Harangi and Virág, Chapter 6 is a joint work with Lippner [19], Chapter 7 is a joint work with Pósfai and Liu [43], and Chapter 8 is a joint work with Hubai and Lovász [14].

Összefoglalás

Nagyméretű gráfok nagyon sok helyen előfordulnak. Ilyenek a biológiai hálózatok, mint az agy; a szilárd testek részecskéi közti kötések gráfja, az internet, illetve a közlekedési, az elektromos, vagy akár a társadalmi hálózatok is. Ezek többségét nemcsak a mérete, de a nehéz hozzáférhetősége miatt is nehéz pontosan megismerni. Arra azonban még így is lehet esélyünk, hogy a számunkra fontos tulajdonságaikat feltérképezzük.

Goldreich és Ron [28] korlátos fokú gráfokra vonatkozó modelljében az alábbi módon definiáljuk a mintavételezést. Veszünk egy r és egy n konstanst, és minden G esetén egyenletes véletlennel kiválasztunk n pontot, és mindnek tekintjük a konstans sugarú környezetét. Egy gráfparaméter *becsülhető*, ha minden $\varepsilon > 0$ -ra van olyan ún. becslő függvényünk, mely minden G gráfra teljesíti, hogy ha bemenetként megkap egy véletlen mintát, akkor ahhoz legfeljebb ε várható hibával a gráf paraméterértékét rendeli. Néhány legegyszerűbb gráfparaméter, amikkel foglalkozhatunk, az expanzió, a legnagyobb független halmaz méretének a gráf csúcsszámához viszonyított aránya, vagy ugyanez lefogó halmazra, vagy párosításra.

A témakörnek egy központi kérdése, hogy a gráfok milyen környezeteloszlásokat adhatnak ki. Ez a kérdés még csoportelméleti sejtésekkel is szoros kapcsolatban áll.

Lokális algoritmusnak nevezzük azokat a függvényeket, melyek valamilyen r-re az r sugarú lehetséges környezetekhez rendelnek valamit. Például lokális algoritmussal független halmazt csinálni azt jelenti, hogy minden csúcsról az r sugarú környezete alapján kell eldönteni, hogy bevesszük-e a halmazba. Ennek egy kiterjesztése, amikor a gráf csúcsaira véletlen számokat sorsolunk, és a döntés függhet a környezeten belülre sorsolt véletlen számoktól. További kiterjesztést jelent, ha kisorsolunk egy közös véletlent is az összes csúcs számára, vagy akár a gráfról adunk meg valami közös információt.

Általában a lokális algoritmusoktól csak közelítő megoldást várunk. Például azt mondjuk, hogy lokális algoritmussal készíthető közel maximális méretű független halmaz, ha minden $\varepsilon > 0$ -ra létezik olyan lokális algoritmus, ami minden G gráfra mindig független halmazt ad, és ennek várható mérete legfeljebb εn-nel kisebb a G-beli maximális független halmaz méreténél.

A lokális algoritmusok szoros kapcsolatban állnak a paraméterbecsléssel. Ha például készíthető lokális algoritmussal közel maximális párosítás, akkor a maximális párosításbeli pontok aránya becsülhető.

A 2. fejezetben mutatunk egy lokális algoritmust a közel maximális folyamra és a közel minimális vágásra. Utána mutatunk két példát, hogy ezek hogyan alkalmazhatóak a gráfok környezeteloszlásainak kérdéseiben. A 3. fejezetben megmutatjuk, hogy a lokális algoritmusoknál bemenetként megadni izomorfia erejéig az egész gráfot is, vagy csak még egy közös véletlent küldeni, ezek ugyanannyira erős kiterjesztések. A 4. fejezetben bebizonyítjuk az algoritmikus eldönthetetlenségét egy a lehetséges köenyezeteloszlásokról szóló kérdéskörnek. Az 5. fejezetben azt az érdekes speciális problémát vizsgáljuk, hogy a 3-reguláris nagykörű gráfokon mekkora független halmaz konstruálható lokális algoritmussal. A 6. fejezetben lokális algoritmusok egyfajta limeszével konstruálunk teljes párosítást nemamenábilis Cayley-gráfokon. A 7. fejezet egy példán keresztük bemutatja, ahogy a fizikusok kutatják a témát, és hogy ez hogyan kapcsolódik a matematikai elmélethez. Végül, a 8. fejezetben mutatunk egy példát arra, hogy ezek az elméletek hogyan kapcsolódnak a matematika más területeihez.

A 2., 3. és 4. fejezet saját egyszerzős cikkeimen alapul [16, 17, 18]. Az 5. fejezet közös munka Gerencsér Balázzsal, Harangi Viktorral és Virág Bálinttal, a 6. fejezet közös munka Lippner Gáborral [19], a 7. fejezet közös munka Pósfai Mártonnal és Yang-Yu Liuval [43], a 8. fejezet pedig közös munka Hubai Tamással és Lovász Lászlóval [14].